ISSN 3035-1863

Journal for the Philosophy of Mathematics

Volume 2
2025

FIRENZE UNIVERSITY PRESS

Journal for the Philosophy of Mathematics

Volume 2 - 2025

Firenze University Press

Journal for the Philosophy of Mathematics

Published by
Firenze University Press University of Florence, Italy
Via Cittadella, 7 - 50144 Florence - Italy
https://riviste.fupress.net/index.php/jpm/index

Contents

Editorial

You are looking at the second issue of the Journal for the Philosophy of Mathematics and the first under my editorship. The issue features seven articles, five of which were invited by the previous editor. All seven underwent a rigorous double-blind refereeing process. We are grateful to those who accepted our invitation and to all the other authors who submitted articles over the past year.

Looking ahead, we hope to increase the proportion of unsolicited submissions. The journal's mission is to encompass the full spectrum of the philosophy of mathematics, publishing philosophical reflections on any aspect of mathematics. This includes foundations, but also pure mathematics beyond foundations, applied mathematics, the history of mathematics, the sociology of mathematics, mathematical practice, mathematical education, and more, provided these areas are meaningfully connected to philosophical questions. We welcome original submissions across all these domains.

This year sees the award of our first Early Career Prize. Open to PhD candidates and scholars within three years of completing their PhD, the prize has been awarded to Davide Sutto at the University of Oslo. We are pleased to publish Davide's article in this volume and to host an online lecture in early 2026 in which he will present his work.

I would like to thank my predecessor, Leon Horsten, for helping to establish the journal and setting it on a successful path. My gratitude also goes to the five Managing Editors for their work this year: Riccardo Bruni, Silvia De Toffoli, Carlo Nicolai, Georg Schiemer, and especially Giorgio Venturi, who has liaised with our publisher, Florence University Press. Thanks also to João Vitor Schmidt for his essential work on the production side. Finally, I wish to thank all the referees who reviewed this year's submissions. This issue stands as a testament to your collective dedication and expertise.

A.C. Paseau
Editor in chief

Journal for the Philosophy of Mathematics

Firenze University Press
www.fupress.com

OPEN ACCESS

Citation: FUJIMOTO, Kentaro (2025). Against Plural Comprehension. *Journal for the Philosophy of Mathematics*. 2: 9-33. doi: 10.36253/jpm-3294

Received: January 28, 2025

Accepted: November 23, 2025

Published: December 30, 2025

ORCiD
KF: 0000-0002-4830-5861

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Against Plural Comprehension

KENTARO FUJIMOTO

School of Mathematics, University of Bristol, UK.
Email: kentaro.fujimoto@bristol.ac.uk

Abstract: Plural primitivism is the idea that plural expressions cannot be dispensed with in favor of singular expressions. Our current standard first-order logic is based on the opposite idea, *singularism*, that plural expressions are eliminable in terms of singular expressions. Hence, plural primitivism suggests replacing first-order logic with what is nowadays called *plural logic*. One prominent axiom of plural logic is the axiom scheme of *plural comprehension* (PCA). This article aims to critically examine the plural primitivist claim of the logicality of PCA.

Keywords: Plural Logic, Plural Comprehension, Second-order Logic, Plurals.

1. Introduction

What I call *plural primitivism* is the idea that plural expressions cannot be dispensed with in favor of singular expressions and should be counted in the irreducible primitive vocabulary of our logic. Our current standard first-order logic is based on the opposite idea, *singularism*, that plural expressions are eliminable in terms of singular expressions: first-order logic only admits singular terms and singular predications about individual objects. Hence, plural primitivism suggests replacing first-order logic with what is nowadays called *plural logic*.[1]

The syntax of plural logic augments that of first-order logic with additional syntactic categories of *plural terms*, including *plural variables* which we will denote by xx, yy, and zz, and *plural quantifiers*, which we will denote by $\forall xx$, $\forall yy$, and $\forall zz$. In addition to them, plural logic has a special binary logical predicate $\prec$ that takes a singular term t in the first argument place and a plural term tt in the second and thereby expresses 't is one of tt'.[2]

[1] This article follows the formalism of plural logic in (Rayo, 2002) and (Florio and Linnebo, 2021), which is called *plural first-order logic* PFO. There are other formalisms that employ different notations and choices of logical vocabulary, but my arguments in this article do not depend on how plural logic is formalized; see (Oliver and Smiley, 2016, Ch. 7) for a comprehensive list of different formalisms of plural logic in the literature.

[2] We may add *non-logical* plural predicates to PFO, which take plural terms (possibly as well as singular terms) as their arguments, and the resulting system is called PFO$^+$ in (Florio and Linnebo, 2021; Rayo, 2002). This article is primarily concerned with plural terms and quantifiers and does not discuss PFO$^+$, but all the criticisms of PFO I will present below equally apply to PFO$^+$.

How is this new formal glossary to be understood? One of the central tenets of plural primitivism is the *ontological innocence* of plural terms and quantifiers. According to plural primitivism, plural terms and quantifiers have distinctive semantic functions which are different from those of singular terms and quantifiers. A plural term is said to *plurally refer* to multiple objects in a different way than a singular term *singularly refers* to a single object: it stands in a one-many referential relation to multiple objects all at once, rather than the ordinary one-one referential relation to a single object, and is alleged to not commit its user to any one-over-many object that somehow comprises those multiple objects, such as the set or class of them. A plural quantifier is said to *plurally quantify* over the first-order domain of discourse, which consists of first-order (singular) objects, and plural primitivism contends that we can thereby quantify over multiplicities of first-order objects without requiring the existence of anything beyond the first-order objects. For example, $\exists xx \forall x (x \prec xx \leftrightarrow x \notin x)$ is interpreted to mean the following English plural construction:

(1) There are some sets such that any one of them is not a member of itself and such that any set that is not a member of itself is one of them.

Plural primitivism contends that since (1) does literally not claim the existence of any set of non-self-membered sets, the formula $\exists xx \forall x (x \prec xx \leftrightarrow x \notin x)$ does not ontologically commit us to the Russell set and, furthermore, is *trivially true*.[3]

One prominent axiom of plural logic is the axiom scheme of *plural comprehension* (PCA henceforth):

$$\exists x \varphi(x) \rightarrow \exists xx \forall x \big(x \prec xx \leftrightarrow \varphi(x)\big), \tag{PCA}$$

where φ is any formula of plural logic without xx free. According to the aforementioned interpretation of the plural glossary, this is interpreted to mean the following statement in English:

(2) If there is something that is φ, then there are some things such that any one of them is φ and such that anything that is φ is one of them,

which neither includes any mention of sets nor commits us to the existence of any set. There is known to be a 'canonical' mutual interpretation between the standard system PFO of plural logic and the system MSOL of monadic second-order logic. In view of this mutual interpretation, PCA corresponds to the second-order axiom (scheme) of impredicative comprehension of (monadic) second-order logic:

$$\exists X \forall x \big(Xx \leftrightarrow \varphi(x)\big), \tag{SCA}$$

where φ is any formula of MSOL without X free. This mutual interpretation suggests the so-called *plural interpretation* of (monadic impredicative) second-order logic, in which second-order variables are interpreted as plural variables (unless their values are not empty). Many advocates of plural primitivism regard PCA, or SCA under the plural interpretation, just as a trivial

[3] Resnik (1988) and Parsons (1990) object that despite appearance, we often need to understand plural terms as referring to sets or similar collection-like objects in order to process sentences containing plural terms. In this article, I proceed on the assumption of the ontological innocence of plural parlance and then aim to show that this assumption renders the plural primitivist case for PCA untenable—or, at least, unconvincing.

or *a priori* logical truth and takes this triviality of PCA as one of the main merits of plural primitivism.[4]

This article aims to critically examine PCA and offer a comprehensive argument that, while it can be true in many contexts and circumstances, the alleged logicality, a priority, and triviality of the truth of the full version of PCA are highly controversial.[5]

To conclude this introductory section, I introduce one notational convention. Plural primitivist parlance is sometimes difficult to express grammatically, unequivocally, and/or idiomatically in English. For example, 'some things are φ' or 'there are some things such that φ' is ambiguous: on the one hand, it may mean that there are two or more things each of which satisfies a singular predicate φ; on the other hand, it may also mean that there are some things that collectively satisfy a plural (collective) predicate φ. Moreover, as Resnik (1988) and Parsons (1990) note, the plural primitivist translation of universal plural quantification is clunky and hard to read. Accordingly, I will use the singular term 'plurality' to mean *many* in the plural primitivist sense—that is, what plural primitivists take plural nouns to denote. For example, to express the collective reading of 'some things are φ', I will use the singular construction 'there is a plurality that is φ'; I will render a universal plural quantification, $\forall xx\varphi(xx)$, as 'every plurality is φ' or 'for every plurality, φ', rather than the official (and awkward) plural primitivist paraphrase, 'it is not the case that there are some things such that it is not the case that φ'. However, the reader should always bear in mind that in such sentences, 'plurality' does not refer to any singular 'one over many' object that somehow 'comprises' several objects, such as a set, class, and platonist universal.

2. PCA as a schema of definitions

Elaborate justifications of PCA are scarce in the literature, perhaps reflecting the aforementioned common plural-primitivist view that its truth is trivial. Nonetheless, scattered remarks in the literature hint at why they think it is, and I will discuss them in turn. The first thought is that PCA is a schema of definitions. For example, Hossack (2014, 526) and Florio and Linnebo (2021, 229) write as follows:

> [U]nlike the other comprehension axioms, the plural axiom does not subserve ontology. Instead it subserves deduction, by underwriting our introduction of new denoting expressions. Given the Theory of Descriptions, [PCA] licenses us, whenever we are given a formula $\phi(u)$, to define 'the $\phi(u)$-ers' accordingly, provided at least one thing satisfies the formula $\phi(u)$. (Hossack, 2014, 526);

> Provided that a condition is well defined and has at least one instance, *of course* the condition can be used to define a plurality of all and only its instances. (Florio and Linnebo, 2021, 229)

[4] Florio and Linnebo (2021) write '[m]any philosophers regard [PCA] as utterly trivial and insubstantial'; we will see several examples of such philosophers shortly.

[5] The same or similar diagnoses are reached elsewhere in the literature via a variety of arguments; e.g., Resnik (1988), Parsons (1990), Hazen (1993), Linnebo (2003), and Hossack (2014), to list a few; Rumfitt (2018) also gave an interesting argument that PCA is inconsistent with a neo-Fregean abstraction principle for ordinals, by which he suggests that at most the Δ^1_1-fragment of PCA can be logically true. Some of my arguments resemble or overlap with theirs, and I will return to them in due course. Hossack (2014) and Florio and Linnebo (2020, 2021) also propose alternatives to PCA: Hossack incorporates stratification into the axiom of comprehension, inspired by Quine's NF; Florio and Linnebo, rather inspired by the Zermelo-Fraenkel set theory, replace comprehension with separation. Florio and Linnebo's theory, called *critical plural logic*, is particularly relevant to the discussion of this article, and I will return to it in due course.

We note that both Hossack (2014) and Florio and Linnebo (2021) deny PCA and intend to characterize their opponents in these quotes.

For notational convenience, let $\imath x K(x)$ formally denote the definite singular description 'the thing (in the domain of discourse) that is K' or 'the K' for short, and $\hat{x}\, K(x)$ formally denote the definite plural description 'the things (in the domain of discourse) that are K'.[6] According to this view, PCA is regarded as a schema of definitions of plural terms $\hat{x}\, K(x)$ for predicates/formulae K of plural logic. In this section, I examine this *definitional* view of PCA.

2.1. PCA *as a schema of naming*

But what does it mean for a term $\hat{x}\, K(x)$ to be *defined* via PCA? We ought to answer this question to assess whether PCA is justifiable as a schema of definitions. Boolos, an arch plural primitivist, expresses the following thought about PCA as a schema of definitions:

> Like the familiar condition: $\exists x \forall y (Ky \leftrightarrow y = x)$ which must be satisfied by a definite singular description 'The K' for its use to be legitimate, there is an analogous condition that must be satisfied by definite plural descriptions. In the simplest case, in which a definite plural description such as 'the present kings of France' is the plural form of a definite singular description, the condition amounts only to there being one object or more to which the corresponding count noun in the singular description applies. . . . Thus like the definite singular description "The K," which has a legitimate use iff the K exists, i.e. iff there is such a thing as the K, "The Ks" has a legitimate use iff the Ks exist, i.e. iff there are such things as the Ks, iff there is at least one K. (Boolos, 1985, 164–5)

It is widely accepted that the existence of a *unique* K justifies the legitimacy of the definite singular description 'the K'. Boolos draws an analogy and contends that the existence of *at least one* K justifies the legitimacy of the definite plural description 'the Ks' in precisely the parallel way. However, this analogy is inappropriate.

The condition $\exists x \forall y (Ky \leftrightarrow y = x)$ for the legitimacy of the singular term 'the K' not only guarantees the non-emptiness of the condition K. It also assures us, by virtue of the initial existential quantifier '$\exists x$', that the range of first-order quantifiers, that is, the domain of all possible referents of singular terms, contains a (unique) entity to be named 'the K'—or any other singular term—that has the desired property K. Hence, the noun phrase 'the K' can be viewed as only *naming* something that has already been presented to us by the holding of the condition $\exists x \forall y (Ky \leftrightarrow y = x)$, and we can legitimately regard the following definitional schema of $\imath x K(x)$ as a *naming* principle:

(3) $\exists! x K(z) \rightarrow \forall z \big(z = \imath x K(x) \ \leftrightarrow \ K(z) \big)$, for all predicates $K(x)$.

Boolos draws an analogy and claims that whenever there are one or more K, then 'the Ks' is a legitimate plural noun phrase *plurally referring to* all and only objects that are K. For him, the following is also a legitimate naming schema:

[6] The term 'definite plural description' is ambiguous in the context of plural logic, since it can be interpreted to mean either the unique plurality that satisfies the condition in question, or the plurality that consists of all and only the things that satisfies the condition. Oliver and Smiley (2016) call the former type of a definite plural description a 'plurally unique description' and the latter type an 'exhaustive description'. Throughout this article, I use 'definite plural description' to mean Oliver and Smiley's exhaustive description.

(4) $\exists x K(z) \to \forall zz\big(zz = \hat{x}\, K(x) \;\leftrightarrow\; \forall z(z \prec zz \leftrightarrow K(z))\big)$;

or, equivalently under the axiom of plural extensionality,

(5) $\exists x K(z) \to \forall z\big(z \prec \hat{x}\, K(x) \;\leftrightarrow\; K(z)\big)$,

from which PCA follows by existential generalization. This (5) is a formalization (in plural logic) of the following principle that Boolos alleged as a truism in the quote above:

(6) If there is some thing that is K, then 'the Ks' is a legitimate term so that anything that is K is among what 'the Ks' plurally refers to, and that anything that is among what 'the Ks' plurally refers to is K.

However, the non-emptiness condition $\exists x Kx$ alone does not assure, unless PCA is presupposed, that a plural variable, such as zz in (4), can have a value to be named $\hat{x}\, K(x)$ that meets the desired property $\forall z\big(z \prec \hat{x}\, K(x) \leftrightarrow K(z)\big)$. In a manner of speaking, the legitimacy condition $\exists x Kx$ in (4) and (5) for a definite plural description $\hat{x}\, K(x)$ achieves only a half of the job that the legitimacy condition $\exists x \forall y(Ky \leftrightarrow x = y)$ in (3) for a definite singular description undertakes, and the other half is only achieved by PCA. This yields a substantial difference. The principle (3) is *conservative*: it adds to no logical truth that has no occurrence of the defined term $\imath x K(x)$, and $\imath x K(x)$ is always eliminable in a deduction of a logical truth including no $\imath x K(x)$. By contrast, (5) is *not* conservative: in particular, they imply the instance of PCA for K, in which $\hat{x}\, K(x)$ does not occur. In my opinion, Boolos's analogy is only a weak analogy and falls short of justifying PCA as a naming principle.

Indeed, the alleged parallelism between (3) and (4) is not acceptable from some foundational standpoints. For example, predicativists deny the definite totality of all subsets of ω and reject the legitimacy of quantification over them. However, they accept many subsets of ω, such as $\{0\}$ and ω itself. Hence, under the assumption of (6), the definite plural description 'the things that are subsets of ω'—or, more simply, 'the subsets of ω'—would count as legitimate definite plural descriptions even for predicativists. But it should not. For, otherwise, although this would not commit predicativists to the existence of any special *object*, such as the powerset of ω, it would still license them to quantify over all subsets of ω via locutions such as 'any one of the subsets of ω' and 'some one of the subsets of ω'. Hence, from a predicativist point of view, (4) and (6) are highly controversial.

2.2. PCA *as a schema of definitions of plural membership*

Another possible understanding of PCA as a schema of definitions is to view (5) as making a definition of the *plural membership relation* $t \prec \hat{x}\, K(x)$. At first glance, one might find it reasonable to say that $K(z)$ 'defines' what it is that $z \prec \hat{x}\, K(x)$ via the biconditional $\forall z(z \prec \hat{x}\, K(x) \leftrightarrow K(z))$.

However, the term 'define' here cannot be understood as an explicit definition in the usual sense that it introduces a mere abbreviation $t \prec \hat{x}\, K(x)$ of $K(t)$. As we have noted in §2.1, the introduction of the term $\hat{x}\, K(x)$ together with its 'definitional' clause $\forall z(z \prec \hat{x}\, K(x) \leftrightarrow K(z))$ yields new logical truths that do not contain $\hat{x}\, K(x)$; namely, the addition of $\hat{x}\, K(x)$ is not conservative. One might still find it reasonable to say that $K(z)$ 'defines' what it is that $z \prec \hat{x}\, K(x)$ in some informal, intuitive sense or might try to give some formal definition of a 'definition' in this sense. However, no matter how the term 'define' would be defined (either

informally or formally), the very idea that $K(z)$ defines what it is that $z \prec \hat{x}\, K(x)$ is vulnerable to the familiar 'vicious circle' argument.

For example, consider Frege's definition of the concept of natural numbers:

$$\forall X\big(\big(0 \in X \wedge \forall z(Xz \to X(z+1))\big) \to Xx\big).$$

Let $\Psi_F(x)$ denote this formula. When we apply the plural interpretation to $\Psi_F(x)$, it is translated into the following predicate:

> It is not the case that there are some things such that 0 is one of them, that if anything is one of them then so is its successor, and that x is not one of them.

Let us call this predicate $\psi_F(x)$. According to the definitional reading of (5) in question, $\psi_F(z)$ defines what it is that $z \prec \hat{x}\psi_F(x)$. Then, whether a given object z is one of $\hat{x}\psi_F(x)$ should be determined by following the definition of $z \prec \hat{x}\psi_F(x)$. However, this is not possible. The membership relation $z \prec \hat{x}\psi_F(x)$ is defined as the satisfaction of ψ_F by z. Hence, to determine whether $z \prec \hat{x}\psi_F(x)$, we have to determine whether $\psi_F(z)$. To determine it, we take an arbitrary plurality yy and then determine whether yy is inductive and whether $z \prec yy$. However, it can be the case that yy happens to coincide with $\hat{x}\psi_F(x)$. Hence, this process involves determining whether $z \prec \hat{x}\psi_F(x)$, which is exactly what we originally wanted to do. By this familiar kind of an argument, (4) should not be read as defining the predicate '$z \prec \hat{x}\, K(x)$': in simple words, this is because the definiens refers to the definiendum in such a definition.[7]

3. PCA as a trivial truth

Some plural primitivists regard PCA as a truism. I have argued that PCA cannot be justified as a consequence of a schema of definitions. Those plural primitivists might instead think the other way around: it is not that PCA is justified by the legitimacy of the definite plural descriptions $\hat{x}\, K(x)$s, but rather that the legitimacy of $\hat{x}\, K(x)$s follows from the self-explanatory truth of PCA. This suggests the most direct justification of PCA, namely, that it is a trivial truth. For example, Lewis and Boolos are among those plural primitivists who regard PCA as such:

> Examples to show the evident triviality of a principle of plural 'comprehension': If there is at least one cat, then there are some things that are all and only the cats. (Regimented ... then there are some things such that, for all x, x is one of them iff x is a cat.) Likewise, if there is at least one set, then there are some things that are all and only the sets. (Lewis, 1991, p. 63)

> [T]he translation of the notorious $\exists X \forall x(Xx \leftrightarrow x$ is not a member of $x)$, where the first-order variables are taken to range over absolutely all sets is "(If there is a set that is not a member of itself, then) there are some sets that are such that each set that is not a member of itself is one of them and each set that is one of them is not a member of itself," as vacuous an assertion about sets as can be made, as desired. (Boolos, 1985, p. 76)

[7] Hossack (2014) makes a similar vicious circle argument against PCA.

Among others Hossack and Uzquiano also call PCA a 'harmless *a priori*' truth (Hossack, 2000, p. 422) and a 'evident triviality' (Uzquiano, 2003, p. 77).[8]

But why is PCA a trivial truth? As Frege (essentially) showed, PCA makes the existence of the least fixed-point of any positive operator a logical truth, which seems to be a highly non-trivial consequence. PCA and its canonical English translation (2) justify (5) and (6), respectively, as naming principles, but, as we have seen, they would license one to make definite plural reference to all and only subsets of ω, which is unacceptable for predicativists. There exist philosophers who seek an alternative to PCA, such as Hossack (2014) and Florio and Linnebo (2021, Ch. 12), and their proposals should not be dismissed by blaming them for missing a trivial truth. It rather seems to me a trivial truth that PCA is not a trivial truth. Having said that, for the sake of the subsequent argument, let us try to examine the alleged triviality of the truth of PCA in more depth.

Why do Boolos, Lewis, and others take PCA to be trivially true? The underlying thought seems to be as follows: any non-empty predicate K *determines* a plurality by prescribing its (plural) membership condition, under which all and only things that satisfy K have that membership. Once such a plurality is determined and given to us, we can name it $\hat{x}\,K(x)$ (or whatever name one likes).

However, a 'vicious circle' argument similar to the one made in §2.2. can be raised against the thought in question, according to which the plural membership of $\hat{x}\,K(x)$ is determined by the predicate $K(z)$. Suppose $K(z)$ is of the form $\forall yy\psi(x, yy)$ with a universal plural quantifier (e.g., $\psi_F(z)$ taken in §2.2.). Take any object a. Since $\psi(a, \hat{x}\,K(x))$ is a substitution instance of a direct sub-formula of K, it appears to be sensible to say that whether $K(a)$ or not is partly determined by whether $\psi(a, \hat{x}\,K(x))$ or not. However, $\psi(z, yy)$ may contain a sub-formula of the form $t \prec \hat{x}\,K(x)$ for various terms t. Hence, whether $\psi(a, \hat{x}\,K(x))$ or not is also partly determined by whether various objects have the membership of $\hat{x}\,K(x)$ or not. We thereby get involved in a vicious circle of determination, in which the determinans refers to the determinatum.[9]

This typical pattern of the vicious circle argument applies to other attempts to legitimize a plurality that appeal to other justificatory relations in which the plurality stands to some predicate K. For example, one might claim that $\hat{x}\,K(x)$ has a definite plural membership *because* K is definite so that it definitely demarcates the domain of discourse into two realms, that is, the things that are K and those that are not. Let $K(x)$ be $\forall yy\psi(x, yy)$ as before. We may, then, ask why K is definite? One sensible answer is that it is *because* $\psi(z, yy)$ is definite no matter what plural referents are assigned to the plural variable yy. Hence, K is definite partly because $\psi(x, \hat{x}\,K(x))$ is definite. We may keep asking why, and, as before, the answer may ultimately be that it is partly *because* $\hat{x}\,K(x)$ has a definite plural membership. One might alternatively say that the definiteness of the plural membership of $\hat{x}\,K(x)$ is *explained by*, or *reduced to*, the definiteness of K. Then, similarly, the definiteness of K is partly explained by, or reduced to, the definiteness of $\psi(z, \hat{x}\,K(x))$ and, ultimately, the definiteness of the plural membership of $\hat{x}\,K(x)$. We get involved in a vicious circle of reasons, explanations, or reductions.

I admit that these are sloppy arguments. For the first vicious circle argument, it can be objected that the sense in which K is said to determine the membership of $\hat{x}\,K(x)$ is different from the sense in which the substitution instance $\psi(a, \hat{x}\,K(x))$ is said to (partly) determine $\forall yy\psi(a, yy)$ (i.e., $K(a)$). Similarly, it can be objected that the sense in which the definiteness

[8] Hossack later changed his view in (Hossack, 2014).

[9] Linnebo (2018) raises essentially the same argument against SCA in terms of grounding.

of $\hat{x}K(x)$ is explained by, or reduced to, the definiteness of $K(x)$ is different from the sense in which the definiteness of $K(x)$ is (partly) explained by, or reduced to, the definiteness of $\psi(x, \hat{x}K(x))$; one can raise a similar objection regarding the sense of 'because'. All these are fair objections: the vicious circle arguments at stake employ the terms like 'determine', 'explain', 'reduce', and 'because' without sufficiently delineating their meanings.

Nevertheless, these vicious circle arguments raise one general concern about justifications of PCA of the kind at stake. Recall that these justifications, as well as that of PCA discussed in §2.2., appeal to a certain justificatory relation in which the plural membership of a plurality xx stands to some predicate K prescribing the (plural) membership condition of xx, under which all and only things that satisfy K have the membership. This amounts to the idea that pluralities are justified as the *extensions* of predicates —or the extensions of what those predicates denote, such as propositional functions, if one does not want to let linguistic expressions in themselves to have extensions. However, didn't we learn the lesson from the foundational debate over mathematics in the early 20th century that this idea does not provide an adequate justification for impredicative comprehension?

Parsons (2002) concisely characterizes Russell's and Weyl's predicativism as the view that 'what are called sets are extensions of concepts' (p. 374); essentially the same view is shared by Poincaré and Frege. As Goldfarb (1989) points out, the need for ramification in Russell's theory does not come from any constructive idea about sets (or 'classes' in Russell's own terms), but from his thought that logic concerns propositional functions rather than sets. A set may be given different specifications belonging to different levels in Russell's ramified hierarchy, but its identity is solely determined by its members; hence, if Russell's theory were a theory of sets, then it would require no ramification. By contrast, identity of a propositional function is not determined by the objects of which it is true, but by way of the manner in which it is 'presented' (in terms of Goldfarb, 1989). Different presentations correspond to different propositional functions, even if they are true of exactly the same objects; in terms of Quine's famous example, 'is an animal with kidneys' and 'is an animal with a heart' present two distinct propositional functions while yielding the same set. For Russell, a set (if it exists) is the extension of a propositional function, which gives a specification of the members of the set, and the membership relation is defined in terms of the satisfaction of the propositional function.[10] Similarly, Weyl denies the realist view that Bernays (1983) later dubbed the 'quasi-combinatorial' concept of sets,[11] and advocates that *infinite* sets can only be justified as extensions of properties that can be constructed ('derived' from 'primitive properties') in a certain manner.[12] Hence, while they do not reject sets outright, both Russell and Weyl view a set as determined by, and auxiliary to, something else that prescribes a condition for a thing to be a member of the set. For Russell, the set membership is defined/determined/explained in terms of the satisfaction of propositional functions; for Weyl, it is defined/determined/explained in terms of the exemplification of those constructible properties. When they are asked why a certain set exists, their answer would be that it is because such and such a propositional function or property exists.

[10] Hence, Russell's theory does not primarily concern sets. He did not take sets as the most fundamental entities of logic, but as something for more practical purposes; for example, Russell says 'the chief purpose which classes serve, and the chief reason which makes them linguistically convenient, is that they provide a method of reducing the order of a propositional function [without affecting the truth or falsehood of its values]' (Russell, 1908, 242).
[11] See (Weyl, 1918, p. 23), for example.
[12] See (Weyl, 1918, §4–§8), for example.

A justification of PCA of the kind at stake goes parallel. It says that a plurality is determined by some predicate K that prescribes its (plural) membership condition; it says that a given object a is (or is not) a plural member of the Ks because a satisfies (or does not satisfy) K. However, a plurality xx could be justified as the extension of a predicate K, only when K is a legitimate/meaningful predicate. Russell's, Weyl's, and other predicativists' criticism of impredicative comprehension concerns the very legitimacy/meaningfulness of impredicative predicates K—or the existence of the propositional function or property that K 'presents' (in terms of Goldfarb, 1989 in his exposition of Russell) or 'expresses' (in terms of Weyl). Hence, merely associating a plurality with a predicate K does not address the predicativists' concern about impredicative comprehension. A natural question, then, is why a justification of the kind at stake succeeds for PCA while the same kind of justification is widely considered inadequate for SCA. Vicious-circle arguments of the familiar sort we have just discussed have driven many philosophers and logicians toward predicativism. Why, then, can only PCA resist them? Even if it can, this is not a trivial matter and would call for a non-trivial argument.

4. The Gödel-Bernays realism

From the discussion in §2.2. and §3, I tentatively conclude that we should not take each instance of PCA,

$$\exists x\varphi(x) \rightarrow \exists xx\forall x\big(x \prec xx \leftrightarrow K(x)\big),$$

to be true by virtue of any reductive, explanatory, or determination relation in which the plural membership $x \prec xx$ stands to the predicate K, and that we cannot justify PCA by regarding pluralities as extensions of predicates. This (tentative) conclusion naturally suggests that a plausible justification of PCA requires that what pluralities exist, and what their plural members are, should be determined independently of us and any linguistic description by us. This view may be called 'realism' about pluralities. Indeed, realism is a standard strategy to justify impredicative definitions in set theory. Gödel famously appealed to realism of sets in his defense of impredicative definitions in set theory. While he acknowledges that the construction or definition of a thing 'can certainly not be based on a totality of things to which the thing to be constructed belongs' (Gödel, 1983, p. 136), he suggests that

> If, however, it is a question of objects that exist independently of our constructions, there is nothing in the least absurd in the existence of totalities containing members which can be described . . . only by reference to this totality. (*ibid*.)

The idea being considered here is 'realist' only in the sense that the extension of the plural membership relation $\prec$, the range of plural (or second-order) quantifiers, and the truth values of plural sentences are objectively determined independently of us; it should not be confused with the idea that there exists an object, independently of us, that can be a single referent of a plural term, and the range of plural quantifiers consists of some such objects, which would collapse plural primitivism into singularism.

However, the crucial problem with this 'realist' conception of plurals is that it completely dissociates plural membership relation $\prec$ from various possible conditions K, whereby the alleged *a priori* connection between the membership of $\hat{x}\,K(x)$ and its description $K(z)$ is totally lost. With the 'realism' of pluralities, a condition or specification, like K, is a mere means of picking out a plurality from the domain of plural quantifiers *if such a plurality belongs to that*

domain. We need an extra argument to guarantee, for each condition K, that all and only objects that are K stands in the independently given relation $\prec$ to *some* plurality. To my understanding, the same line of consideration led Bernays to propose the 'quasi-combinatorial' conception of sets in justification of impredicative definitions in set theory, according to which a set is viewed as

> the result of infinitely many independent acts deciding for each number whether it should be included or excluded. (Bernays, 1983, p. 260)

Namely, the set $\{z \in x \mid K(z)\}$ exists as 'the result of infinitely many independent acts of' including objects in the set only when K is true of them (and excluding objects from it only when K is false of them).[13]

Can we adapt the Gödel-Bernays 'realist' argument for a justification of PCA as a logical axiom? On the one hand, it does not seem to be straightforward to adapt the Gödelian realism for a justification of PCA. Pluralities are actually not such objects that exist, or do not exist, in the ordinary sense according to the plural primitivist thesis of the ontological innocence of plural terms. Hence, the desired adaptation requires to spell out what it means that pluralities and the totality of them objectively exist independently of us.

On the other hand, Bernays's notion of 'quasi-combinatorial' definitions of sets can be relatively straightforwardly adapted to definitions of pluralities. A plurality can then be understood as the outcome of possibly infinitely many independent acts of picking objects and deciding whether to include them in or exclude them from that plurality.

In my opinion, however, the Bernaysian 'quasi-combinatorial' conception of pluralities hardly justifies PCA *as a logical axiom*. It requires infinitary acts of selecting things. Set theory is a descriptive science of a specific subject matter, that is, sets and their universe. Metaphorically speaking, even though we humans cannot perform infinitely many acts, God can do it in His creation of the universe of sets, and that universe may well contain the results of such infinitary acts by God. Logic, by contrast, is not a descriptive science of any specific subject matter; rather, it is supposed to be topic-neutral and universally applicable, studying valid inferences that preserve truth in all possible circumstances. I find no compelling reason to assume that the domains of discourse in all those circumstances involve the results of such infinitary acts: God may decide not to exercise His ability to perform infinitary acts in creating some of those circumstances. Furthermore, to me, logic is the study of reasoning by us, not by God, and thus logic should not postulate an axiom that presupposes something that we can never perform even in ideal circumstances.

After all, PCA under the Gödel-Bernays realist conception of pluralities is apparently no more logical than the axioms of separation in set theory and can hardly be called a logical axiom.[14]

[13] Linnebo (2003, §IV) argues that PCA requires this Gödel-Bernays realism, and maintains this view also in (Florio and Linnebo, 2021), where they write 'To say that plural comprehension is permissible on a condition φ is to say that we may reason quasi-combinatorially about all the φs' (p. 229).

[14] Florio and Linnebo (2021, Ch. 10) present an argument for the quasi-combinatorial conception of pluralities, which relies on the assumption of the *traversability* of pluralities. However, formulating this assumption requires infinite disjunctions of arbitrarily large cardinality and thus, in my view, covertly commits us to infinitary acts (of articulating infinitely long conditions). Their argument provides a basis for *critical plural logic*, their theory of *circumscribed* pluralities, in which pluralities are conceived of as extensionally definite, modally rigid, and traversable. However, they themselves do not regard it as a logic in the sense at issue in this article; see (Florio and Linnebo, 2021, Ch. 12.7).

5. Semantics

There might be other, more plausible 'realist' justifications of PCA as a logical truth than the Gödel-Bernays realism, although I am not currently aware of any. In this section, rather than exploring such realist justifications further, I turn to a crucial presupposition of the Bernaysian 'quasi-combinatorial' conception, namely, the *definiteness* of impredicative plural formulae. An infinitary selection of the members of a plurality could not be carried out without the definiteness of a condition K according to which the members are selected. More generally, the definiteness of K seems to be a necessary condition for any justification of the definiteness of the definite plural description 'the Ks'. Furthermore, if we succeed in justifying the definiteness of K without being committed to a vicious circle, then we can simply legitimize 'the Ks' as a term plurally referring to the extension of K. However, the claim that every formula/predicate is definite has never been unanimously accepted in the history of the philosophy of mathematics; for example, intuitionism, strict finitism, and constructivism reject it. This is why Gödel appealed to realism about sets and their totality to guarantee the definiteness of conditions K, in justifying impredicative definitions in set theory. In what follows, I present a problem with the alleged definiteness of impredicative plural formulae from a semantic point of view.

5.1. Semantic Determinacy and Commitment

Let $\mathfrak{M}$ be a first-order structure for a first-order language $\mathcal{L}$. The extensional meanings of $\mathcal{L}$-expressions, such as the truth values of $\mathcal{L}$-sentences and the referents of closed $\mathcal{L}$-terms, are determined solely by $\mathfrak{M}$. Now, suppose we augment $\mathcal{L}$ with a new first-order predicate P and constant c, and let us call the thus extended first-order language $\mathcal{L}'$. The referents of closed $\mathcal{L}'$-terms including c is not determined solely by $\mathfrak{M}$, and nor is the truth value of $\mathcal{L}'$-sentences including P. This is because $\mathfrak{M}$ tells us nothing about P and c. We may describe this situation as the *semantic determinacy* of $\mathcal{L}$ in $\mathfrak{M}$ and the *semantic indeterminacy* of $\mathcal{L}'$ in $\mathfrak{M}$. To make $\mathcal{L}'$ *semantically determinate*, we typically augment $\mathfrak{M}$ with extra semantic information, namely, fixed interpretations of P and c.

For a more relevant example, let $\mathcal{L}^2$ be the second-order extension of $\mathcal{L}$ which augments $\mathcal{L}$ with second-order variables and quantifiers. Under Henkin semantics, $\mathcal{L}^2$ is not semantically determinate in $\mathfrak{M}$. To make it semantically determinate, we need to augment $\mathfrak{M}$ with a set of subsets of the first-order domain of $\mathfrak{M}$ as the range of second-order quantifiers. By contrast, $\mathcal{L}^2$ *is* semantically determinate in $\mathfrak{M}$ alone under the standard ('full') semantics of second-order logic, in which second-order quantifiers are automatically interpreted to range over absolutely all subsets of the first-order domain of $\mathfrak{M}$: no new information beyond the specification of that first-order domain is required.

Now, under Henkin semantics, we may call a second-order language *semantically further-committal* (relative to first-order logic), in the sense that its semantic determinacy requires further *semantic* information beyond that supplied by the semantic interpretation of its first-order part. By contrast, under the standard semantics, a second-order language is not semantically further-committal: a first-order model-theoretic structure solely determines the semantics of all its expressions and, in particular, make all its predicates definite.

Semantic commitment, in this sense, is closely related to ontological commitment, but the two notions are independent. Under Henkin semantics, in order for $\mathcal{L}^2$-expressions to receive definite extensional meaning, further objects beyond those supplied by $\mathfrak{M}$—namely, a set of

subsets of the domain of $\mathfrak{M}$ and its members—are required, and thus we may say that $\mathcal{L}^2$ is ontologically further committed to these objects. Hence, a second-order language under Henkin semantics is both semantically and ontologically further-committal. By contrast, a second-order language under the standard semantics is only ontologically further-committal and not semantically further-committal. The converse does not necessarily hold either: further semantic commitment does not entail further ontological commitment. For example, Florio and Linnebo (2016) propose the *plurality-based Henkin semantics* of plural logic, in which a plural language is interpreted by a first-order structure augmented with a *super-plurality* over the first-order domain—that is, a plurality of pluralities of first-order objects—as a range of plural quantifiers: a plural quantifier is then interpreted as ranging over the plural members of that super-plurality.[15] A superplurality is a plurality of pluralities of individual objects and alleged to incur no ontological commitment beyond those individual objects. Hence, under the plurality-based Henkin semantics, plural expressions are considered to be not ontologically further committal, while they are semantically further committed to that super-plurality.[16]

5.2. *The Maximum Domain Thesis*

Is a language of plural logic semantically further-committal or not? Most plural primitivists seem to think that it isn't, and even take this semantic not-further-committalness as one of the main virtues of plural primitivism. For those plural primitivists, the range of plural quantifiers is automatically and uniquely fixed once the semantic interpretation of the first-order vocabulary is fixed. Let us call this view the *unique domain thesis*. This can be compared to the semantic treatment of the logical relation of identity: its extension is uniquely fixed once the semantic interpretation of the (non-logical) first-order vocabulary is fixed.

But what should such a uniquely determined range be like? Recall that the standard semantics of second-order logic renders a second-order language semantically not-further-committal because it automatically interprets second-order quantifiers to range over *absolutely all* subsets of the first-order domain. Many plural primitivists draw an analogy from the standard semantics of second-order logic, and take plural quantifiers as ranging over *absolutely all* pluralities of first-order objects. Let us call this idea the *maximum domain thesis*. The unique domain thesis does not imply the maximum domain thesis, but the unique domain thesis without the maximum domain thesis appears to be unnatural.[17] Indeed, the maximum domain thesis is quite common among plural primitivists.[18]

5.3. *Interdependence with sets*

In this sub-section, however, I argue that the maximum domain thesis is in tension with the alleged logicality of the plural vocabulary and thus of PCA.

[15] For discussions of superpluralities, see also (Hazen, 1997), (Oliver and Smiley, 2005, 2016), (Rayo, 2006), (Linnebo and Nicolas, 2008), and (Florio and Linnebo, 2021).

[16] Florio and Linnebo (2016) offer another option for the range of plural quantifiers in a plurality-based Henkin model, that is a plural property: then, plural quantifiers are interpreted as ranging over the pluralities of first-order objects that satisfy that plural property.

[17] I am aware of only one natural alternative, namely, the *predicativist* domain thesis that plural quantifiers range over all and only pluralities describable by formulae with no plural quantifiers; however, this alternative clearly fails to justify PCA anyway.

[18] According to Florio and Linnebo (2016), 'nearly all writers who have embraced plural logic on the plurality-based semantics ascribe to this system metalogical properties which presuppose that the semantics is standard rather than Henkin' (p. 566). Note that Florio and Linnebo (2016) themselves do not even embrace the unique domain thesis.

5.3.1. Etchemendy's argument

We begin with Etchemendy's (1990) well known argument. He contends that the Tarskian definition of logical consequence is inappropriate for second-order logic under the standard semantics, because it renders many distinctively mathematical statements, such as the continuum hypothesis (CH), either logically true or logically false. It is well known that, under the standard semantics, there are formulae $N(X)$ and $R(X)$ of some language of MSOL that pin down (categorically axiomatize) the domains $\mathbb{N}$ and $\mathbb{R}$ of the standard models of arithmetic and real ordered field (up to isomorphism), respectively. Using these, one can construct a sentence Γ such that Γ is logically true in MSOL iff CH is true and that Γ is logically false in MSOL iff CH is false.[19] Hence, if MSOL under the standard semantics is a genuine logic, then CH is either logically true or logically false.

One might think that a similar objection applies to plural logic under the maximum-domain thesis, but the situation is not so simple. To illustrate this, for each formula Φ in MSOL, let $\Phi^{(p)}$ denote the canonical translation of Φ in PFO; conversely, for each formula φ in PFO, let $\varphi^{(1)}$ denote the canonical translation of φ in MSOL. Then, $\Gamma^{(p)}$ is a natural candidate for a PFO-sentence to express CH. However, the original Γ is equivalent to CH in MSOL under the (set-based) standard semantics because second-order quantifiers are interpreted as ranging over sets. Given that the plural quantifiers in $\Gamma^{(p)}$ do not range over sets, $\Gamma^{(p)}$ need not be equivalent to Γ: without further assumptions, it can be the case that there is a plurality yy such that no set Y is coextensive with yy; it can equally be the case that there is a set Y such that no plurality yy is coextensive with Y.

Of course, Γ and $\Gamma^{(p)}$ are equivalent under the assumption that sets necessarily coincide with pluralities—namely, that, necessarily, for every set X there is a plurality of the members of X, and that, necessarily, for every plurality xx, there is a set of the plural members of xx—, but this assumption appears to be unmotivated from the plural primitivist point of view. First, if pluralities necessitate coextensive sets, then they are naturally taken to carry ontological commitment to sets. Second, plural primitivists aim to offer a set-free alternative logic that is still as strong as second-order logic, so they would want to avoid positing such a tight logical/metaphysical connection between pluralities and sets. Third, the assumption entails that only set-sized pluralities exist, thereby precluding the plural interpretation of proper classes.

Furthermore, even when xx and X are coextensive, $N^{(p)}(xx)$ need not be equivalent to $N(X)$. Again, without further assumptions, it can be the case that there is a plurality yy that is not coextensive with any set Y but witnesses the ill-foundedness of xx; it can equally be the case that there is a set Y that is not coextensive with any plurality yy but witnesses the ill-foundedness of X. Hence, either or both of $N^{(p)}(xx)$ and $N(X)$ may fail to pin down $\mathbb{N}$. Similarly, either or both of $R^{(p)}(xx)$ and $R(X)$ may fail to pin down $\mathbb{R}$.

Even though $\Gamma^{(p)}$ may not be equivalent to CH, and even though $\Gamma^{(p)}$ may not be about the genuine $\aleph_0$ and $\aleph$, it still feels like a mathematical statement. However, given that $\Gamma^{(p)}$ lacks any logical connection with CH, it is less clear whether it should be treated as an extra-logical statement. Indeed, many apparently distinctively mathematical statements—for example, that any group of order 361 $(= 19 \times 19)$ is abelian—are logical truths even in first-order logic. Since

[19] See (Etchemendy, 1990, Ch. 9, note 11). In MSOL, we actually need an extra non-logical assumption—for example, an assumption that enables us to code an ordered pair of any two objects—to construct such a sentence Γ. Jané (2005) presents essentially the same line of argument against second-order logic under the standard semantics.

plural primitivists aim to drastically strengthen logic via plural vocabulary, they might be prepared to treat more statements of more mathematical flavour, including $\Gamma^{(p)}$, as logically true (or logically false).

In what follows, I supplement Etchemendy's argument by presenting two further cases in which, under reasonable assumptions, certain sentences of plural logic stand in a more direct interdependence with distinctively mathematical statements.

5.3.2. Case 1

Let $\mathcal{L}^2_{\mathbb{N}}$ and $\mathcal{L}^p_{\mathbb{N}}$ denote the languages of second-order and plural arithmetic, respectively, whose non-logical vocabulary is exactly that of the first-order language $\mathcal{L}_{\mathbb{N}}$ of arithmetic. We denote the standard model of arithmetic by $\mathfrak{N}$, which is a structure for both $\mathcal{L}^2_{\mathbb{N}}$ (under the standard semantics) and $\mathcal{L}^p_{\mathbb{N}}$ (under the maximum domain thesis).

We define the classes Π^p_n and Σ^p_n ($n \in \mathbb{N}$) of $\mathcal{L}^p_{\mathbb{N}}$-formulae by the obvious analogy with the classes Π^1_n and Σ^1_n ($n \in \mathbb{N}$) of $\mathcal{L}^2_{\mathbb{N}}$-formulae: every $\mathcal{L}^p_{\mathbb{N}}$-formula without plural quantifiers is $\Pi^p_0 = \Sigma^p_0$; if $\varphi(xx)$ is Σ^p_n, then $\forall xx\varphi(xx)$ is Π^p_{n+1}; if $\varphi(xx)$ is Π^p_n, then $\exists xx\varphi(xx)$ is Σ^p_{n+1}; then, $\Phi^{(p)}$ is equivalent (in PFO) to a Σ^p_n-formula for any Σ^1_n-formula Φ, and $\varphi^{(1)}$ is equivalent (in MSOL) to a Σ^1_n-formula for any Σ^p_n-formula φ.

According to the maximum domain thesis, $\mathfrak{N} \models \forall xx\varphi(xx)$ holds if and only if $\mathfrak{N} \models \varphi(xx)$ for absolutely all pluralities xx of natural numbers. But what pluralities are included in that range? Take any *set* X of natural numbers, which belongs outside the domain $\mathbb{N}$ of $\mathfrak{N}$. Consider the definite plural description 'the things that are members of the set X'. This description is *predicative*—in the sense that it contains no plural quantifiers—in a language that can express the membership relation and whose domain of discourse contains X. Hence, it is much more harmless and reasonable to treat it as a legitimate definite plural description than impredicative ones. Hence, although X belongs outside $\mathbb{N}$, it exists anyway, and it seems plausible to me, under the maximum domain thesis, that the plurality of the members of X also exists. More generally, the following assumption appears to be quite plausible:

(7) For every set X, there is a plurality of the members of X.

Let $\mathfrak{A}$ be any first-order structure with domain A. Under the assumption of (7), the range of plural quantifiers in $\mathfrak{A}$ subsumes the powerset of A.

We next assume the following:

(8) The least infinite ordinal ω in the universe of sets is isomorphic to the domain $\mathbb{N}$ of the standard model of arithmetic.[20]

Under the assumptions (7) and (8), if a Σ^1_1-sentence Φ is true in the universe of sets, then $\Phi^{(p)}$ is also true in $\mathfrak{N}$. For example, let Ψ be a Σ^1_1-sentence expressing that there is a set of natural number that codes a (countable) model of ZFC plus one inaccessible cardinal. If the universe of sets is a model of ZFC and contains two inaccessible cardinals, then Ψ is true in the universe of sets (by the reflection principle and Löwenheim-Skolem theorem), from which it follows that $\mathfrak{N} \models \Psi^{(p)}$. By contraposition, if $\mathfrak{N} \not\models \Psi^{(p)}$, then there is at most one inaccessible cardinal in the universe M of sets (if the universe of sets is a model of ZFC).

[20] If we work within a set-theoretic meta-theory, (8) means that the model of object set theory is an ω-model of set theory.

Furthermore, many plural primitivists (e.g., Hossack, 2000; McKay, 2006, Ch. 6; Oliver and Smiley, 2016, Ch. 13.2) hold the following:

(9) Plural logic can pin down (categorically axiomatize) the standard model $\mathfrak{N}$ of arithmetic;

they often take (9) to be one of the main advantages of plural logic over first-order logic. Let Ξ be a sentence of plural logic that pins down $\mathfrak{N}$. Under the assumption of (9), the existence of a countable model of ZFC plus one inaccessible cardinal implies that $\Xi \wedge \Psi^{(p)}$ is logically true. By contraposition, if $\Xi \wedge \Psi^{(p)}$ is logically false, then the universe contains at most one inaccessible cardinal (again if the universe of sets is a model of ZFC).

This may be taken to indicate that which pluralities exist, or which sentences are logically true in plural logic, depends on which sets exist, or conversely that which sets exist depends on which pluralities exist or which sentences are logically true in plural logic. That said, Ψ and $\Psi^{(p)}$ (or $\Xi \wedge \Psi^{(p)}$) are still not equivalent, and it may be that $\Psi^{(p)}$ is true and Ψ is false. In the next subsubsection, I present a stronger case.

5.3.3. Case 2

We are presently comparing plural quantification over a first-order domain with singular quantification over subsets of the same domain, allowing for the possibility that the two are substantially different. To proceed with the current argument in a formally precise way, we need to fix a meta-theory in which the semantics for both types of quantification are defined. In what follows, we work within a sufficiently rich set-theoretic meta-theory and assume that plural logic (under the maximum domain thesis) receives the set-based Henkin semantics there. We denote the (genuine) least infinite ordinal ω in the meta-theory by $\mathbb{N}$, regarded as the domain of the standard model $\mathfrak{N}$ of arithmetic, and let $Pl(\mathbb{N})$ be some set of subsets of $\mathbb{N}$ serving as the range of plural quantifiers in $\mathfrak{N}$.

First, it seems reasonable to make the following assumption:

(10) The $\mathcal{L}^p_{\mathbb{N}}$-formula asserting the well-foundedness of a sub-plurality of $\mathbb{N}$—which is a Π^p_1-formula—is always correct in $\mathfrak{N}$;

that is, $(\mathbb{N}, Pl(\mathbb{N}) \cup \{\emptyset\})$ is a β-model of second-order arithmetic.

Next, we fix a model $\mathfrak{M} = (M, E)$ of set theory that is defined within the meta-theory. The domain M of $\mathfrak{M}$ will be regarded as 'the universe of sets' at the object level. I make two assumptions about $\mathfrak{M}$:

(11) $\mathfrak{M}$ is a well-founded model of some moderately strong theory T extending, say, the Kripke-Platek set theory KPω plus Σ_1-separation.
(12) The height of $\mathfrak{M}$ is sufficiently large so that the order-type of the ordinals in $\mathfrak{M}$ is greater than or equal to ω_1^L in the meta-theory;[21]

By (11), $\mathfrak{M}$ is isomorphic to some transitive model of T in the meta-theory (assuming that the meta-theory can prove Mostowski's collapsing theorem). Hence, in particular, $\omega^{\mathfrak{M}}$ is isomorphic to $\mathbb{N}$, and thus we will identify $\omega^{\mathfrak{M}}$ and $\mathbb{N}$ in what follows. Moreover, T is rich enough so that the constructible universe $L^{\mathfrak{M}}$ can be defined in its model $\mathfrak{M}$ and that many basic facts about

[21] We can replace ω_1^L with a smaller ordinal, such as, the least stable ordinal.

it are true in $\mathfrak{M}$, such as the Shoenfield absoluteness. Both assumptions, (11) and (12), do not concern pluralities; rather, they are purely about the universe of sets (at the object level).

I will show that the truth values of any Σ^1_2-sentence Φ (in $\mathfrak{M}$) and its $\mathcal{L}^p_{\mathbb{N}}$-translation $\Phi^{(p)}$ (in $\mathfrak{N}$) coincide: that is, $(\mathbb{N}, Pl(\mathbb{N})) \models \Phi^{(p)}$ iff $(\mathbb{N}, \mathcal{P}(\mathbb{N})^{\mathfrak{M}}) \models \Phi$.

First, because PCA is a logical axiom and thus true in $\mathfrak{N}$, the $\mathcal{L}^p_{\mathbb{N}}$-translation $\mathsf{Z}_2^{(p)}$ of full second-order arithmetic Z_2 is true in $\mathfrak{N}$. Hence, in $\mathfrak{N}$, the constructible hierarhy (up to a certain level) can be coded in $\mathcal{L}^p_{\mathbb{N}}$. By (10), the codes of constructible sets in $\mathbb{N}$ are correct, and there is an ordinal ρ at the meta-level such that the union of the constructible sets that are coded in $\mathcal{L}^p_{\mathbb{N}}$ in $\mathfrak{N}$ coincides with L_ρ in the meta-theory. Furthermore, the Shoenfield absoluteness holds in $\mathfrak{N}$ in terms of these codes of constructible sets: any Σ^1_2-sentence Φ is true in in L_ρ, iff $\Phi^{(p)}$ is true in $\mathfrak{N}$.[22]

Second, by (11), the constructible hierarchy $L^{\mathfrak{M}}$ in $\mathfrak{M}$ coincides with the genuine one at the meta-level up to a certain level, say, L_τ.

Third, by (7), (10), and (12), the supremum of the order-types of well-orderings of $\mathbb{N}$ in $\mathfrak{N}$ is $\geq \omega_1^L$, and thus $\rho \geq \omega_1^L$. Moreover, it is also assumed that $\tau \geq \omega_1^L$ by (12).

Finally, take any Σ^1_2-sentence Φ. By the Shoenfield absoluteness in $\mathfrak{N}$, $\Phi^{(p)}$ is true in $\mathfrak{N}$, iff Φ is true in L_ρ. Since $\rho, \tau \geq \omega_1^L$, Φ is true in L_ρ, iff Φ is true in L_τ. By the Shoenfield absoluteness in $\mathfrak{M}$, Φ is true in L_τ, iff Φ is true in $\mathfrak{M}$.

There are many distinctively mathematical Σ^1_2-statements. For example, the existence of a countable transitive model of any recursive set theory, such as ZFC plus two inaccessible cardinals, is a Σ^1_2-statement. The axiom of Σ^0_n-determinacy ($n \in \mathbb{N}$) is also Σ^1_2.[23] These Σ^1_2-sentences may be undecidable in T: the existence of a countable transitive model of ZFC plus two inaccessible cardinals is independent of ZFC; Σ^0_n-determinacy for large n is independent of some weak theory T that meets the condition (11).[24] However, the truth of these statements in the universe M of sets are equivalent to the truth of its canonical $\mathcal{L}^p_{\mathbb{N}}$-translation in $\mathfrak{N}$. If we further assume (9), then they are either logically true or logically false. This mirrors Etchemendy's set up, in which a statement independent of the standard set theory ZFC (i.e., CH) comes out either logically true or logically false.

It might be objected that the assumptions (11) and (12) are ad hoc and that we need not accept them. However, they are assumptions purely about the universe of sets, asserting some desirable, or at least reasonable, property of the universe, and they do not concern pluralities. We can also cook up various different assumptions that yield similar consequences. The import of what we have discussed so far is that under the assumption of (7), certain additional assumptions purely about sets—e.g. (11) and (12)—and certain additional assumptions purely about pluralities—e.g. (10)—may together yield a strong interdependence between sets and pluralities.

5.3.4. One possible way out — the plurality-based Henkin semantics

These examples suggest that, under the maximum domain thesis, some logical truths in plural logic are interdependent with the ontology of sets and the structure of their universe. This

[22] See (Simpson, 2009, Ch. VII.4) for the details of the coding of constructible hierarchy in second-order arithmetic as well as the proof of Shoenfield absoluteness in terms of that coding.
[23] See (Simpson, 2009, V.8) for the definition of Σ^0_n-determinacy in the language of second-order arithmetic.
[24] If the meta-theory decides these Σ^1_2-sentences, then $\mathfrak{M}$ also satisfy them by the condition (12).

consequence poses a challenge to the plural primitivists' claim that plural logic is a genuine logic.

One possible way to avoid this consequence is to adopt a semantics that is semantically *further-committal* but still ontologically not-further-committal, such as the aforementioned plurality-based Henkin semantics by Florio and Linnebo (2016). This might be an effective plural primitivist rejoinder to the argument I have just presented, but is not without problems. First, the notion of super-plurality (or plural property) is controversial. Second, it requires justification of the existence of a super-plurality (or plural property) that is closed under impredicative definitions of pluralities of first-order objects. Third, to make PCA a logical axiom under their semantics, all the models taken into account by the semantics need to satisfy PCA, but this requirement must be justified. Indeed, Florio and Linnebo give up PCA after all and propose what they call *critical plural logic* (Florio and Linnebo, 2020, 2021) in which PCA is severely restricted.[25]

6. Indescribable pluralities

The other possible plural primitivist rejoinder to the argument in §5.3. is to deny (7) and hold that the existence, or non-existence, of a plurality coextensive with a set X is completely independent of the existence, or non-existence, of X or, more generally, independent of any object of any kind that is not included in the first-order domain. However, then, my concern is that, without any connection with other collection-like objects, our ordinary understanding of plural reference and quantification seems to be hopelessly *uninformative* about to which pluralities we can plurally refer.

We apparently know the truth condition of, and can determine the truth value of, 'some boys are chatting at the corner', independently of any set or any object other than those mentioned in the sentence. However, in everyday use of plurals like in this example, we only take *finite* pluralities into account, but we are currently concerned with more abstract and theoretical settings in which infinite pluralities need to be considered. We human beings can, in principle, point to or explicitly list finitely many objects, but we cannot do the same for infinitely many.

To plurally refer to infinitely many objects, we usually use definite plural descriptions, such as 'the prime numbers'. However, the truth value of a sentence may rely on the existence of pluralities that cannot be described in a given language, even by impredicative predicates. Let us see one example. For $\mathcal{L}^p_{\mathbb{N}}$-formulae $\varphi(xx)$ and $\theta(x)$, let $\varphi(\{x|\theta(x)\})$ denote the result of replacing each occurrence of $t \prec xx$ in φ with $\theta(t)$ (for each term t). This $\varphi(\{x|\theta(x)\})$ expresses that φ is true of the θs. If $\varphi(\{x|\theta(x)\})$ is true and if $\hat{x}\theta(x)$ is considered a legitimate definite plural noun phrase, then $\exists xx\varphi(xx)$ should be evaluated to be true. However, suppose, in contrast, that $\varphi(\{x|\theta(x)\})$ is false for any plural arithmetical formula θ (even if θ is an impredicative formula). On the one hand, Lévy (1965) showed that it is consistent relative to ZFC that there is a Π^1_2-formula $\Phi(X)$ of second-order arithmetic that admits no projective uniformization; hence, it follows that $\exists X\Phi(X)$ can be true *without* $\Phi(\{x|\Theta(x)\})$ being true for any second-order arithmetical formula Θ. On the other hand, Addison Jr (1959) showed that every second-order arithmetical formula admits a projective uniformization, if $V = L$, and thus $\Phi(\{x|\Theta(x)\})$ is true for some second-order formula Θ whenever $\exists X\Phi(X)$ is true. These results suggest that both $\exists xx\varphi(xx)$ and $\neg\exists xx\varphi(xx)$ may be true when $\varphi(\{x|\theta(x)\})$ is false for any plural

[25] See also footnotes 9 and 17.

arithmetical formula θ; indeed, we can transform Addison Jr's and Lévy's models into models of plural arithmetic in which these are the cases. Hence, both $\exists xx\varphi(xx)$ and $\neg\exists xx\varphi(xx)$ are coherent possibilities, but their truth values rely on the existence (or non-existence) of a plurality that cannot be described by any description.

We are often able to evaluate $\exists xx\varphi(xx)$ to be true by resorting to other objects than those in a given domain of discourse, such as sets of natural numbers. Furthermore, Lévy's (or Addison Jr's) proof provides us with a fairly clear idea of at least one situation (concerning not only natural numbers but also higher-order sets) in which $\varphi(xx)$ is true (false, resp.) for a plurality xx that is coextensive with some set.

However, we are now supposing that pluralities are completely independent of any object outside $\mathbb{N}$, and that all the language-independent factors to determine the truth values of $\mathcal{L}^p_{\mathbb{N}}$-sentences are supposed to be provided by a semantic interpretation of the first-order language $\mathcal{L}_{\mathbb{N}}$ of arithmetic. Hence, $\exists xx\varphi(xx)$ needs to be either true or false even if there is nothing but natural numbers.[26] But, then, what is the fact that is delineated by $\mathfrak{N}$ alone but still determines whether there is some plurality xx such that $\varphi(xx)$? What would a situation be like in which no objects other than the natural numbers exist, yet there exists such an absolutely indescribable, highly complex plurality of natural numbers? I have no idea, and I can hardly imagine that anyone else does.

Our linguistic intuition is of no help in answering these questions. Our ordinary understanding of plurals tells us too little about infinite pluralities that cannot be described within the language we speak—that is, $\mathcal{L}^p_{\mathbb{N}}$, in the case under consideration. Plural primitivists provide a translation of a plural quantifier of PFO into English—that is, $\exists xx\varphi$ is translated into 'there are some things that are φ'—, but this does not add to my understanding of when there are some things that are φ and when there are not. Plural primitivists employ idiosyncratic locutions and say, for instance, that plural quantifiers 'plurally quantify over' the first-order objects, but this informs me of nothing about what infinite plurality can be a value of a plural variable.[27] Our understanding of plurals alone appears to give us no idea of what it is like that there is a plurality, conceived as completely independent of any description or object (set, in particular), that satisfies Lévy's formula.

Taking all this into account, my view is that if pluralities are completely independent of any object outside $\mathbb{N}$, then there is no fact of the matter that determines whether there exists a plurality satisfying Lévy's formula φ. Hence, nothing among those language-independent factors fixed by $\mathfrak{N}$ alone settles whether there is a plurality that satisfies φ. In particular, if pluralities are completely independent of any object outside $\mathbb{N}$, there is no determinate range of plural quantifiers that determines whether $\exists xx\varphi(xx)$ is true. This, in turn, casts a doubt on the definiteness of impredicative $\mathcal{L}^p_{\mathbb{N}}$-formulae in $\mathbb{N}$. Finally, if impredicative $\mathcal{L}^p_{\mathbb{N}}$-formulae are not definite, then the truth of impredicative instances of PCA (at least in $\mathfrak{N}$) is open to doubt. After all, taking pluralities to be completely independent of any objects outside $\mathbb{N}$ would undermine the plausibility of PCA.

[26] Furthermore, we can extend Lévy's theorem to higher-order set theory: if there is a transitive model M of ZFC and the existence of a strongly inaccessible cardinal κ, then we can construct a transitive model $N \supset M$ of ZFC such that κ remains strongly inaccessible in N and that there is a second-order *set-theoretic* formula $\Phi(X)$ such that $V_\kappa \models \exists X\Phi(X)$ but $V_\kappa \not\models \Phi(Y)$ for all ordinal definable subsets of V_κ in N. Hence, there is a model of plural set theory in which $\exists xx\varphi(xx)$ is true but $\varphi(yy)$ is false for any yy that is describable in any higher-order set theory.

[27] Jané (2005, §10) argues that 'assuming that we understand [plural quantification] well enough for everyday purposes is not a ground for believing that it can support canonical second-order consequence', and I fully agree with him.

It might be objected that while our ordinary understanding of the English word 'set' also tell us little about infinite sets, mathematicians nonetheless have a substantial understanding of them. Mathematician's understanding of infinite sets are largely based on the current set theory as a branch of mathematics, which can hardly be called a theorization of our everyday concept of sets. The (set-based) standard semantics of second-order logic, from which plural primitivists draw an analogy and insight in claiming the semantic determinacy and not-further-committalness of plural sentences, is based on such a mathematical meta-theoretic understanding of sets. Hence, one could try to invent a similarly sophisticated theory of infinite pluralities and base their semantics of plurals on it; Florio and Linnebo's critical plural logic (2020) might be a good candidate for such a theory. Our plural primitivist might thereby object that plural logic under the 'standard' semantics with such a theory of pluralities as the meta-theory fares no worse than second-order logic under the standard semantics.

However, first of all, to me, second-order logic under the standard semantics is not a logic. My argument against plural logic in the last section equally (and more straightforwardly) applies to second-order logic under the standard semantics. It is not neutral to the ontology of sets and is highly dependent on the background set theory; hence, for example, Vänäänen (2001, p. 504) concludes that it is just a 'major fragment of [set theory]', and I believe many logicians and philosophers nowadays share the same view, e.g., Koellner (2010).

Second, regardless of my own view of second-order logic, such a 'sophisticated' theory of pluralities likely requires an axiom equally and similarly controversial as PCA. Second-order logic under the standard semantics renders SCA logically valid because the background set theory postulates the axioms of powerset and separation, the latter of which has the same (or even worse) impredicative character as SCA. Similarly, to make PCA valid, the background theory of pluralities would likely need to postulate axioms of a similar impredicative character; for instance, Florio and Linnebo's (2020; 2021) critical plural logic postulate equally impredicative axioms that correspond to the axioms of powerset and separation.

Third, the new conception of pluralities that would be brought to us by such a 'sophisticated' theory of infinite pluralities might well be quite different from, and foreign to, our everyday conception of plurals. Such a theory and its conception of pluralities might turn out to be no more logical or ontologically innocent than the current set theory and its conception of sets; if so, plural logic would be a 'major fragment of' such a non-logical and/or ontologically committal theory.

My own view is this. The question of what constitutes the range of plural quantifiers is the question of what can be *plurally* referred to; otherwise, the plural primitivist notion of ontologically innocent plural reference would give little support to the ontological innocence of plural quantification. Semantics connects a language and the world, and reference is a semantic relation. The world is independent of languages, and objects exist independently of languages. It is part of the job of a semantic interpretation to supply all the language-independent factors needed to determine the referents and truth values of expressions in the language it interprets. It would be no mystery that the domain of discourse a semantic interpretation specifies contains something to which no noun phrase in the language can refer (under that fixed interpretation). However, if plural expressions are semantically not-further-committal, then all the language-independent facts that are relevant to their referents or truth values must be fully given by a semantic interpretation of the first-order vocabulary, which only contains the information as to what first-order objects exist and of which first-order objects each first-order predicate is

true (assuming for simplicity that the vocabulary contains no function or constant symbols). To me, this seems to indicate that there is no fact of the matter, in the circumstance fixed by the semantic interpretation, about plural references that cannot be described by the first-order vocabulary. Hence, in my opinion, while the predicative plural comprehension axiom, in which the condition K in PCA is restricted to be a plural formula with no plural quantifiers, might possibly be a logical axiom (for those who accept the plural vocabulary as logical), the full impredicative PCA cannot.

7. Issues on topic-neutrality

In this section, I give another argument against the logicality of PCA from a different perspective. Namely, I will argue that the assumption of the logicality of PCA has negative consequences upon the topic neutrality of plural logic. My argument in this section does not rely on any strong semantic assumption regarding plural logic. It only requires that the semantics of plural logic is sound for two-sorted first-order logic (where plurals are viewed as the second sort), such as Takeuti's (1987) system **BC**. Hence, one may work with plural logic under the maximum domain thesis, adopt Florio and Linnebo's plurality-based Henkin semantics, or can regard plural logic as a purely syntactic deductive system.

It is widely thought that logic ought to be *topic-neutral* and *universally applicable* to any topic and subject in the same uniform way. There seems, however, nothing that accommodates and applies to absolutely every theorization of every subject with every conception of the subject matter; if this were a requirement for logic, then even classical logic would fall short of logic because it is incompatible with arithmetic with the intuitionistic conception of natural numbers. In my opinion, topic-neutrality is a matter of degree after all, and I think that whether something is logic ultimately should not be judged solely on the basis of which topics it can cover. Nonetheless, if something claimed to be logic covers too limited a range of topics, this is still a negative sign for its logicality. Therefore, if plural logic is a genuine logic, then it should be topic-neutral to a decent extent.

Restriction of impredicative comprehension abounds in mathematical logic. Typical examples are predicativism; when applied to arithmetic, predicativism results in the Weylian predicative arithmetic or Russelian ramified analysis (or its transfinite extension by Kreisel, Feferman, and Schütte). Finitism gives another example: the second-order system RCA_0^* of arithmetic is defined as $\mathsf{I}\Delta_0(\mathsf{exp})$ plus the second-order axiom of induction and the restriction of SCA to Δ_1^0-formulae: the idea behind the RCA_0^* is that only elementarily recursively definable pluralities are admissible in mathematics from a certain strong finitist point of view.

If PCA is logical, any restriction of SCA is no less 'illogical' under the plural interpretation of second-order quantifiers than the suppression of, say, the identity axioms from first-order logic. Now, the second-order axiom IND of induction,

$$\forall X\big(X0 \wedge \forall n(Xn \to Xn+1) \to \forall n Xn\big),$$

is often considered constitutive of the concept of natural number. For example, Dummett (1994, p. 337) advocates that '[i]t is part of the concept of natural number ... that induction with respect to any well-defined property is a ground for asserting all natural numbers to have that property'; Lavine (1994, p. 231, n. 24) also contends that 'part of what it is to define a property of natural numbers is to be willing to extend mathematical induction to it'. Whilst both Dummett and Lavine adopts the property interpretation of second-order

quantifiers here, the same conclusion applies to IND under any other interpretation on the same ground: that is, any multiplicity of natural numbers in any interpretation of the multiplicity, whether it is understood as a property, set, of plurality of natural numbers, has the least element. Hence, any system of plural arithmetic naturally postulates IND on this ground. However, then, all sub-systems of Z_2 are deprived of serious mathematical and/or foundational significance under the plural interpretation of second-order quantifiers, as 'illogical', which renders much of traditional proof theory insignificant and worthless; thereby, the degree of universal applicability and topic-neutrality of plural logic is considerably reduced.

A possible rejoinder from our plural primitivist might be that the systems investigated in proof theory should be understood not as systems in plural logic but as systems in two-sorted first-order (singular) logic whose subject matter is constituted by natural numbers *and* 'ones over many' over natural numbers of a certain kind, such as sets or (platonist) properties of natural numbers.

However, second-order arithmetic under the plural interpretation and that under, say, the set interpretation share the same first-order part anyway. Hence, if plural logic is a genuine logic, then all the *first-order* consequences of Z_2 are 'logical consequences' of Robinson Arithmetic Q and IND (under the plural interpretation). There has been proposed a plethora of purely first-order theories of arithmetic from various foundational points of view that are proper extensions of Q and (interpretable in) proper sub-systems of Z_2.[28] All these theories only concern the common part of the plural interpretation and the set interpretation of second-order arithmetic, and the difference of the two interpretations is irrelevant to them. However, they all fall far below the first-order part of Z_2. Hence, if plural logic with PCA is a genuine logic, and if arithmetical induction is constitutive of the concept of natural numbers, then these theories are all rendered as 'illogical' and insignificant, which still trivializes much of proof theory. Many plural primitivists are perhaps ready to bite the bullet and accept such a large-scale debunking of the traditional proof theory, but it is surely an unattractive and unpleasing option for many philosophers and mathematicians.[29]

Moreover, the 'two-tiered' approach under consideration is faced with a further difficulty when applied to second-order set theory (also known as class theory). What Z_2 is to arithmetic is what the Morse-Kelley theory MK is to set theory, and the study of subsystems of MK has been rapidly developed in recent years from various foundational and philosophical perspectives.[30] What classes are has been one of the central questions in the philosophy of set theory, and the set interpretation is no longer possible for classes due to the existence of proper classes. One of the major merits of plural primitivism is alleged to be that the plural interpretation offers an answer to this question: namely, quantification over classes is plural quantification over sets. However, to maintain the foundational significance and value of the study of subsystems of

[28] Among such theories are the Kreisel-Feferman-Strahm theories of unfolding (Feferman and Strahm, 2000), the Turing-Feferman theories of transfinite progressions of consistency statements or reflection principles (Feferman, 1962; Turing, 1939), first-order theories of generalized inductive definitions and their variants, Feferman's theories of reflective closures (Feferman, 1991), and various axiomatic theories of truth (see Halbach, 2010).

[29] Hazen (1993) also considers the 'two-tiered' view under consideration here: he pointed out that it puts plural primitivists in 'the anomalous position of holding that someone (the predicativist) who accepts part of second-order logic is ontologically committed to more than someone who accepts all of it!'. Although he seems to share essentially the same worry with me, his contention here (sometimes called 'Hazen's puzzle') is already preempted by the plural primitivist rejoinder under consideration. According to the rejoinder, what predicativists accept is a theory of natural numbers *and* sets, while what plural primitivists accept is a theory purely about natural numbers; the former is not 'part of' the latter.

[30] See (Jäger, 2009), (Jäger and Krähenbühl, 2010), (Fujimoto, 2012, 2023), (Sato, 2014, 2015), (Gitman and Hamkins, 2016), and (Gitman et al., 2020), for example.

MK, the 'two-tiered' approach requires classes to be given a different interpretation than the plural one, whereby plural primitivism loses one of its alleged major merits.

8. Summary

Let me summarize what I have argued. First, PCA cannot be justified as a schema of definitions (§2). Second, PCA is hardly a trivial, *a priori*, self-explanatory truth (§3). Third, the orthodox Gödel-Bernays 'realist' justification of impredicative definitions in set theory cannot be employed in justification of PCA as a logical truth (§4). Fourth, one of the central tenets of plural primitivism, the semantic determinacy and not-further-committalness of plurals, suggests the maximum domain thesis, which makes plural logic with PCA dependent on ontology (§5 and §6). Fifth, PCA severely reduces the topic-neutrality of plural logic from the viewpoint of the current practice of logic. Hence, I conclude that PCA is not a logical axiom.

However, even if my conclusion is accepted, it might be argued that the alleged ontological innocence of PCA would still be a significant merit even as a *non-logical* principle; for example, if Z_2 under the plural interpretation is not ontologically committed to anything beyond natural numbers, then it fares better than Z_2 under the set interpretation, in terms of ontological parsimony. I conclude this article with a brief comment on this issue without in-depth discussion.

Even as a non-logical principle, the truth of PCA must be justified whenever it is postulated. Hence, the question is whether such a justification can be ontologically innocent and less metaphysically laden than a justification for SCA. My argument so far seems to indicate that it cannot. Regardless of whether PCA is taken as a logical axiom, it can be justified neither as a schema of definitions nor as a trivial truth: my arguments against these two types of justifications still stand stand regardless of whether PCA is logical. Realism about pluralities might provide a justification for PCA as a non-logical principle. However, as I argued in §6, without some connection to objects outside the domain of a first-order structure, the truth of PCA is not guaranteed.[31] And if justification of PCA appeals to such a connection to extra objects, then it is no longer ontologically innocent. Thus, even as a non-logical principle, PCA does not appear to fare significantly better than SCA in terms of ontological parsimony.

Acknowledgements. I am grateful to the anonymous referees for their helpful comments and constructive suggestions.

References

Addison Jr, J. W. (1959). Some consequences of the axiom of constructibility. *Fundamenta Mathematicae*, 46:337–357. DOI: https://www.doi.org/10.4064/fm-46-3-337-357.

Bernays, P. (1983). On platonism in mathematics. In Benacerraf, P. and Putnam, H., editors, *Philosophy of Mathematics*, pages 258–271. Cambridge University Press, Cambridge.

Boolos, G. (1985). Reading the Begriffsschrift. *Mind*, 94:331–344. Reprinted in Boolos (1998), 155–70. Citation is to reprint. DOI: https://doi.org/10.1093/mind/XCIV.375.331.

Boolos, G. (1998). *Logic, Logic, and Logic*. Harvard University Press, Cambridge, Massachusetts.

Dummett, M. (1994). Reply to Wright. In McGuinness, B. F. and Oliveri, G., editors, *The Philosophy of Michael Dummett*, pages 329–338. Kluwer, Dordrecht. DOI: https://doi.org/10.1007/978-94-015-8336-7_21.

[31] Hence, in my view, PCA is false when the first-order domain contains absolutely every object.

Etchemendy, J. (1990). *The Concept of Logical Consequence*. Harvard University Press, Cambridge, MA.

Feferman, S. (1962). Transfinite recursive progressions of axiomatic theories. *The Journal of Symbolic Logic*, 27:259–316. DOI: https://doi.org/10.2307/2964649.

Feferman, S. (1991). Reflecting on incompleteness. *The Journal of Symbolic Logic*, 56:1–49. DOI: https://doi.org/10.2307/2274902.

Feferman, S. and Strahm, T. (2000). The unfolding of non-finitist arithmetic. *Annals of Pure and Applied Logic*, 104:75–96. DOI: https://doi.org/10.1016/S0168-0072(00)00008-7.

Florio, S. and Linnebo, Ø. (2016). On the innocence and determinacy of plural quantification. *Noûs*, 50(3):565–583. DOI: https://doi.org/10.1111/nous.12091.

Florio, S. and Linnebo, Ø. (2020). Critical plural logic. *Philosophica Mathematica*, 28(3):172–203. DOI: https://doi.org/10.1093/philmat/nkaa020.

Florio, S. and Linnebo, Ø. (2021). *The Many and the One*. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/oso/9780198791522.001.0001.

Fujimoto, K. (2012). Classes and truths in set theory. *Annals of Pure and Applied Logic*, 163:1484–1523. DOI: https://doi.org/10.1016/j.apal.2011.12.006.

Fujimoto, K. (2023). A few more dissimilarities between second-order arithmetic and class theory. *Archive for Mathematical Logic*, 62:147–206. DOI: https://doi.org/10.1007/s00153-022-00829-3.

Gitman, V. and Hamkins, J. (2016). Open determinacy for class games. In Caicede, A. E., Cummings, J., Koellner, P., and Larson, P., editors, *Foundations of Mathematics, Logic at Harvard, Essays in Honor of Hugh Woodin's 60th Birthday*, pages 121–144. American Matematical Society.

Gitman, V., Hamkins, J. D., Holy, P., Schlicht, P., and Williams, K. J. (2020). The exact strength of the class forcing theorem. *The Journal of Symbolic Logic*, 85(3):869–905. DOI: https://doi.org/10.1017/jsl.2019.89.

Gödel, K. (1983). Russell's mathematical logic. In Benacerraf, P. and Putnam, H., editors, *Philosophy of Mathematics*, pages 447–469. Cambridge University Press, Cambridge.

Goldfarb, W. (1989). Russell's reasons for ramification. In Savage, C. W. and Anderson, C. A., editors, *Rereading Russell: Essays in Bertrand Rusell's Metaphysics and Epistemology*, pages 24–40. University of Minnesota Press, Minneapolis.

Halbach, V. (2010). *Axiomatic Theories of Truth*. Cambridge University Press, Cambridge. DOI: https://doi.org/10.1017/CBO9781139696586.

Hazen, A. (1993). Against pluralism. *Australasian Journal of Philosophy*, 71(2):132–144. DOI: https://doi.org/10.1080/00048409312345142.

Hazen, A. (1997). Relations in Lewis's framework without atoms. *Analysis*, 57(4):243–248. DOI: https://doi.org/10.1093/analys/57.4.243.

Hossack, K. (2000). Plurals and complexes. *The British Journal for the Philosophy of Science*, 51(3):411–443. DOI: https://doi.org/10.1093/bjps/51.3.411.

Hossack, K. (2014). Sets and plural comprehension. *Journal of Philosophical Logic*, 43:517–539. DOI: https://doi.org/10.1007/s10992-013-9278-2.

Jäger, G. (2009). Full operational set theory with unbounded existential quantification and power set. *Annals of Pure and Applied Logic*, 160:33–52. DOI: https://doi.org/10.1016/j.apal.2009.01.010.

Jäger, G. and Krähenbühl, J. (2010). Σ^1_1 choice in a theory of sets and classes. In Schindler, R., editor, *Ways of Proof Theory*, pages 283–314. Ontos Verlag, Frankfurt. DOI: https://doi.org/10.1515/9783110324907.283.

Jané, I. (2005). Higher-order logic reconsidered. In Shapiro, S., editor, *The Oxford Handbook of Philosophy of Mathematics and Logic*, pages 781–810. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/oxfordhb/9780195325928.003.0026.

Koellner, P. (2010). Strong logics of first and second order. *The Bulletin of Symbolic Logic*, 16:1–36. DOI: https://doi.org/10.2178/bsl/1264433796.

Lavine, S. (1994). *Understanding the Infinite*. Harvard University Press, Cambridge, Massachusetts.

Lévy, A. (1965). Definability in axiomatic set theory I. In Bar-Hillel, Y., editor, *Logic, Methodology and Philosophy of Science: Proceedings of the 1964 International Congress*, pages 127–151. North-Holland, Amsterdam.

Lewis, D. (1991). *Parts of Classes*. Basil Blackwell, Oxford.

Linnebo, Ø. (2003). Plural quantification exposed. *Noûs*, 37:71–92. DOI: https://doi.org/10.1111/1468-0068.00429.

Linnebo, Ø. (2018). On the permissibility of impredicative comprehension. In Rivera, I. F. and Leech, J., editors, *Being Necessary: Themes of Ontology and Modality from the Work of Bob Hale*, pages 170–187. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/oso/9780198792161.003.0009.

Linnebo, Ø. and Nicolas, D. (2008). Superplurals in English. *Analysis*, 68(3):186–97. DOI: https://doi.org/10.1093/analys/68.3.186.

McKay, T. J. (2006). *Plural Predication*. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/acprof:oso/9780199278145.001.0001.

Oliver, A. and Smiley, T. (2005). Plural descriptions and many-valued functions. *Mind*, 114(456):1039–1068. DOI: https://doi.org/10.1093/mind/fzi1039.

Oliver, A. and Smiley, T. (2016). *Plural Logic*. Oxford University Press, Oxford, second edition. DOI: https://doi.org/10.1093/acprof:oso/9780198744382.001.0001.

Parsons, C. (1990). The structuralist view of mathematical objects. *Synthese*, 84(3):303–346. DOI: https://doi.org/10.1007/BF00485186.

Parsons, C. (2002). Realism and the debate on impredicativity, 1917–1944. In Sieg, W., Sommer, R., and Talcott, C., editors, *Reflection on the Foundations of Mathematics: Essays in Honor of Solomon Feferman*, pages 372–389. Association for Symbolic Logic. DOI: https://doi.org/10.1017/9781316755983.018.

Rayo, A. (2002). Word and objects. *Noûs*, 36(3):436–464. DOI: https://doi.org/10.1111/1468-0068.00379.

Rayo, A. (2006). Beyond plurals. In Rayo, A. and Uzquiano, G., editors, *Absolute Generality*, pages 220–254. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/oso/9780199276424.003.0009.

Resnik, M. D. (1988). Second-order logic still wild. *The Journal of Philosophy*, 85(2):75–87. DOI: https://doi.org/10.2307/2026993.

Rumfitt, I. (2018). Neo-Fregeanism and the burali-forti paradox. In Rivera, I. F. and Leech, J., editors, *Being Necessary: Themes of Ontology and Modality from the Work of Bob Hale*, pages 188–223. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/oso/9780198792161.003.0010.

Russell, B. (1908). Mathematical logic as based on the theory of types. *American Journal of Mathematics*, 30(3):222–262.

Sato, K. (2014). Relative predicativity and dependent recursion in second-order set theory and higher-order theories. *The Journal of Symbolic Logic*, 79:712–732. DOI: https://doi.org/10.1017/jsl.2014.28.

Sato, K. (2015). Full and hat inductive definitions are equivalent in *NBG*. *Archive for Mathematical Logic*, 54:75–112. DOI: https://doi.org/10.1007/s00153-014-0403-x.

Simpson, S. G. (2009). *Subsystems of Second Order Arithmetic*. Cambridge University Press, Cambridge. DOI: https://doi.org/10.1017/CBO9780511581007.

Takeuti, G. (1987). *Proof Theory*. North-Holland, Amsterdam, second edition.

Turing, A. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, 45:161–228. DOI: https://doi.org/10.1112/plms/s2-45.1.161.

Uzquiano, G. (2003). Plural quantification and classes. *Philosophia Mathematica*, 11:67–81. DOI: https://doi.org/10.1093/philmat/11.1.67.

Vänäänen, J. (2001). Second-order logic and foundation of mathematics. *The Bulletin of Symbolic Logic*, 7:504–520. DOI: https://doi.org/10.2307/2687796.

Weyl, H. (1918). *The Continuum: A Critial Examination of the Foundation of Analysis*. The Thomas Jefferson University Press, Kirksville, Missouri. Translated by Stephen Pollard and Thomas Bole.

Journal for the Philosophy of Mathematics

Firenze University Press
www.fupress.com

Structures in Arbitrary Object Theory

LEON HORSTEN

Department of Philosophy, University of Konstanz, Germany.
Email: leon.horsten@uni-konstanz.de

OPEN ACCESS

Citation: HORSTEN, Leon (2025). Structures in Arbitrary Object Theory. *Journal for the Philosophy of Mathematics*. 2: 35-44. doi: 10.36253/jpm-3292

Received: January 28, 2025

Accepted: September 22, 2025

Published: December 30, 2025

ORCiD
LH: 0000-0003-3610-9318

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Abstract: This article critically compares two approaches that seek to explain *ante rem* mathematical structures in terms of arbitrary objects. The first is a theory that was developed by Kit Fine, and the second is a view that was more recently developed by me. I make a plea for a synthesis between the two approaches.

Keywords: Mathematical structuralism, arbitrary object theory, permutation problem.

1. Introduction

Two kinds of mathematical structures can be distinguished: rigid structures, and non-rigid structures. This division is defined in terms of a natural notion of symmetry on a structure. A rigid structure is a structure that admits non-trivial automorphisms, i.e., non-trivial isomorphisms from the structure to itself. A non-rigid structure does not have non-trivial automorphisms.

We will be concerned with the nature of rigid and non-rigid mathematical structures, explicated in an *ante rem* manner[1] in terms of arbitrary objects. The idea of connecting *ante rem* structuralism with the theory of arbitrary objects was first explored by Fine.[2] The present article intends to be a contribution to a discussion between Kit Fine and me about these matters. We have engaged in an exchange of constructive criticism concerning our respective proposals.[3] In this paper I argue that our mutual critiques point in the direction of a *synthesis* of the two approaches that might satisfy both parties.

To some extent, this will be a tale of book reviews. There are book reviews, and then there are *book reviews*. This article is to a significant extent about a couple of examples of the latter: (Burgess, 1999), (Burgess, 2008) and especially (Fine, 2022). We will explore how one might improve on the account of mathematical structures in (Horsten, 2019b) and (Horsten, 2019a), in the light of objections and friendly suggestions by Fine in his book review (Fine, 2022).

[1] The *locus classicus* for the discussion of *ante rem* structuralism is (Shapiro, 1997).
[2] See (Fine, 1998, Section VI).
[3] I have criticised Fine's account in (Horsten, 2019b, Section 8.4.3), and in (Horsten, 2019a, Section 6.1).

I will discuss the problem of describing the nature of rigid and non-rigid mathematical structures on the basis of a particularly simple example of each: the *natural number structure* on the one hand, and the *complete two-element graph* on the other hand. But the philosophical conclusions drawn from these examples are supposed to generalise to all rigid and non-rigid structures.

The structure of this article is as follows. I start with a brief revisit of the main ingredients of Fine's metaphysical theory of arbitrary objects, which I will call **F-theory**, and of my alternative to it, which I will call **H-theory**. I then move on to the way in which Fine and I employ arbitrary object theory to give a philosophical account of *ante rem* mathematical structures. Special attention is given to Fine's claim that my theory cannot do justice to the uniqueness of subject matter of categorical mathematical theories. Also, close attention is given to the metaphysical treatment of non-rigid mathematical structures.

2. Two theories of arbitrary objects

The basic idea of arbitrary object theory is that beside the particular objects that we are all familiar with (chairs, people, individual atoms,...), there are also *arbitrary objects*. An arbitrary object is an object that can take particular objects as values, but is not numerically identical to any particular object. The best case for the existence of such a quaint sort of entities lies, in my view, in its applications in metaphysics. It has been argued, for instance, that random variables can best be seen as arbitrary objects. In this paper, we explore the view that mathematical structures can best be viewed as arbitrary objects.

There are two competing metaphysical theories of arbitrary objects: the one pioneered by Fine in (Fine, 1985) and further developed in later publications (*F-theory*), and the theory proposed in (Horsten, 2019b) and applied in later publications such as (Horsten, 2019a) (*H-theory*).

The bare bones of F-theory are as follows.[4] Arbitrary objects are entities that have *values*. Here 'taking a value' is a primitive metaphysical concept, which is not to be confused with functional application (even though arbitrary objects can be *modeled* as functions.) For any collection of ordinary (*particular*) objects S, there is an *independent arbitrary* S (a_S) such that for each $o \in S$, the arbitrary object a_S takes the value o (and a_S takes no other values). The set S then forms the *value range* of a_S. For instance, if $S = \mathbb{N}$, then a_S is an independent arbitrary natural number. But there are also arbitrary objects that *depend on* other arbitrary objects. For instance, there is a *dependent* a'_S such that, whenever a_S takes the value n, a'_S takes the value $2 \times n$. Thus the dependence relation is an irreducible conceptual component of Fine's theory of arbitrary objects.

In (Horsten, 2019b), an alternative metaphysical account of arbitrary objects is proposed. According to H-theory, arbitrary objects are objects that can be in *states* that make up a *state space*. Thus an arbitrary number can be in the state of being the particular number 29. There are states s such that more than one arbitrary object can simultaneously be in state s. Thus there is a state such that an arbitrary natural number a_1 "is" the number 29 while another arbitrary object a_2 "is" the number 23.

These two metaphysical accounts are closely related to each other.[5] Fine's notion of taking a value corresponds to the H-theory's notion of being in a state, and Fine's related notion of value space corresponds to the H-theory's related notion of state space. But there is no primitive notion of dependence in the H-theory. Rather, in H-theory dependence is a defined notion

[4] See (Fine, 1985, chapter 1) for an introduction to F-theory.

[5] Indeed, Fine conjectures that *mathematically*, F-theory and H-theory are equivalent (Fine, 2022, p. 616–617).

(Horsten, 2019b, section 9.4): it is a matter of correlation of values across states. This then is a fundamental metaphysical difference between the two accounts.

In H-theory, a notion of possibility plays a role. H-theory takes the notion of possibility that is at play to be a peculiar modality (*afthairetic modality*). The reason for this is that it does not seem to make much sense to ask what the *actual* state is that a given arbitrary number is in: so it is a notion of modality without an accompanying notion of actuality. Fine, however, is sceptical about this modality. Moreover, he believes that it is important that his account of arbitrary objects is regarded as a *non-modal* account (Fine, 2022, p. 604):

> "[...] is the underlying theory of the kind of object involved in such statements itself modal? [...] [M]y own view is that it is not. [...] we should dispense with this peculiar modality and talk explicitly about the values that x and y take in different possible situations."

It is not clear that the disagreement between the two accounts on this issue is substantial. It is not an uncommon view that any satisfactory maximally expressive theory of metaphysical possibility should directly quantify over possible worlds—this happens, for instance, when possible worlds semantics for a modal language is explained in a metalanguage. (My metaphysical theory does quantify directly over 'afthairetically possible situations', of course.) Moreover, Fine would presumably agree that it does not make much sense to ask whether there is any situation in the *actual world* where a_S takes the value 21.[6]

At any rate, even if Fine would in the final analysis prefer to avoid mentioning situations at all in his theory of arbitrary objects, modality arguably nonetheless implicitly plays a role in Fine's dependence relation. The conditional that is used to express dependence relations ('if...then...') is clearly not the material conditional. Rather, it is an indicative conditional that is to semantically be understood in modal terms.

H-theory can be seen as an *abstractionist* account of arbitrary objects. It holds that systems of arbitrary objects *supervene* on their behaviour in states. On the one hand, this means that it suffices for the specification of a system $\langle a_1, \ldots, a_n \rangle$ of arbitrary objects to determine what the value of each a_i (for $i \leq n$) is on a set S of states (taking the arbitrary objects a_i to be undefined outside of S). This is in effect a *comprehension principle* for arbitrary objects. On the other hand, it implies a principle of *ontological economy* for arbitrary objects (Horsten, 2019b, p. 45):[7]

> Suppose that a and b are arbitrary objects. Then $a = b$ if and only if for every situation s, the value that a takes in s is identical to the value that b takes in s.

This is in effect a *principle of extensionality* for arbitrary objects.

Even though Fine accepts a principle that is close to the aforementioned extensionality principle, he sees matters differently. For Fine, the dependence relation is primitive and does not supervene on the values that arbitrary objects take. He holds that there is an *initial* arbitrary natural number—let us call it a.[8] Then there also is an arbitrary natural number b that *depends* on a and that takes the value 13 whenever a takes the value 11 and *vice versa*; and moreover there is an arbitrary natural number c that depends on b that takes the value 13 whenever arbitrary

[6] Observe, however, that it seems perfectly reasonable to ask, after reading the above quote by Fine: "But which values to x and y *actually* take?"
[7] Fine accepts a closely related principle: see (Fine, 1985, p. 34).
[8] Cf. infra, p. 39.

natural number b takes the value 11 and *vice versa*. Then a and c always take the same values. However, c (indirectly) depends on a, but not *vice versa*.

At the informal level there is a connection between (mathematical) structures and arbitrariness (Burgess, 2008, p. 403):

> [Both (*in rebus* structuralism and *ante rem* structuralism)] may be taken to start with the idea that 'the real numbers' means 'the arbitrary complete ordered field' and then diverge over the interpretation of 'arbitrary'. On the majority view, 'the arbitrary F' denotes nothing by itself: 'The arbitrary F is a G' just means All Fs are Gs'. On the minority view (as in Fine, 1985) the arbitrary F is a specific F but an extraordinary one in that it has no properties not shared by all Fs (though it is distinguished from ordinary Fs by the 'meta-property'[9] of being arbitrary).

Both F-theory and H-theory take the 'minority' approach: they aim to develop the thought that arbitrary object theories can be fruitfully employed in the construction of theories of *ante rem* mathematical structures.[10]

3. Fine's theory of mathematical structure

Fine's arbitrary objects account of mathematical structures is based on (Fine, 1985). It is sketched in a condensed form in (Fine, 1998, section VI), and was later slightly modified in (Fine, 2022). In this section, we concentrate on the presentation of the account in (Fine, 1998, section VI).

The process of obtaining a structure in terms of arbitrary objects consists of three stages. When we apply this to the case of the natural number structure, the account is as follows (Fine, 1998, p. 630):

Stage I We consider the *independent* arbitrary object N which, for every ω-sequence of particular objects $\underline{N}$, has $\underline{N}$ as one of its values (and has no other values). This arbitrary object N is called the *prototype* of the natural number structure.

Stage II We consider the sequence of *dependent* arbitrary objects

$$a_0^N, a_0^N, \ldots, a_i^N, \ldots$$

such that for every natural number i, a_i^N is the arbitrary object that takes for every value $\underline{N}$, of N, the i-th element of $\underline{N}$ as a value (and takes no other values).

Stage III The natural number structure is then defined as the following structure (in the classical set theoretic sense of the word):

$$\mathbb{N}_F = \langle \{a_0^N, a_1^N, \ldots, a_i^N, \ldots\}, < \rangle,$$

where $<$ is the ordering naturally induced by the orderings on the values of N.

According to Fine's account, arithmetic is then about one unique subject matter (Fine, 2022, p. 606–608). This subject matter is not an arbitrary object, but a *particular* object: it is a particular structure in the classical *set theoretic* sense of the word, i.e., a particular *set*. Thus Fine's account

[9] This relates to Fine's distinction between *generic* and *classical* conditions (Fine, 1985, p. 14), which I leave aside here.

[10] One might be sceptical about *ante rem* structures in general, of course, but in this article we do not enter into the discussion about the relative advantages and disadvantages of *ante rem* structuralism and *in rebus* structuralism.

of the natural number structure is a version of *set theoretic structuralism*. Indeed, we immediately see that:

Proposition 3.1. $\mathbb{N}_F$ *is an* ω*-sequence.*

So there is a possible situation in which $\mathbb{N}_F$ takes itself as its value.

Fine distinguishes between representational and non-representational kinds of objects: "Following Hallett . . . , we shall say that an account of the types of some kind is *representational* if each type of the given type is of that type." (Fine, 1998, p. 623) Thus, for instance, von Neumann's definition of cardinal numbers is representational: the cardinal number of a two-element set contains two elements. On the other hand, Frege's definition of cardinal numbers is not: the cardinal number two is an infinite class. The previous proposition then says that Fine's definition of natural number structure is representational.

I have argued[11] that the F-theory suffers from a *permutation problem* in the sense of (Hellman, 2006, p. 546). The objection goes as follows. Consider the structure $\mathbb{N}'_F$, which is just like $\mathbb{N}_F$, except that the places of a_0^N and a_1^N in the ordering $<$ are reversed. Clearly this also is an ω-sequence. According to Fine's account, $\mathbb{N}'_F$ is not, however, the natural number structure. Nonetheless, for familiar Benacerrafian reasons, it is hard to give convincing reasons for why $\mathbb{N}_F$, rather than $\mathbb{N}'_F$ is the 'real' natural number structure.

Even if this worry is dismissed, a closely related argument raises a complication for F-theory. Fine holds that there are not just one but many *independent* arbitrary real numbers (Fine, 2022, p. 609),[12] so presumably also many independent arbitrary natural numbers, and many natural number prototypes. Each of the latter gives rise to an ω-sequence that is a candidate for being the subject matter of arithmetic: will the privileged ω-sequence please rise? Fine argues that among the independent arbitrary reals, only one is the *initial* independent arbitrary real—the Ur-arbitrary real, we might say (Fine, 2022, p. 609). Likewise, presumably, there is a unique initial natural number prototype, and the subject matter of arithmetic is the ω-sequence that *this* naturally gives rise to. Thus the notion of being 'initial' appears to be an additional primitive notion in Fine's theory of arbitrary objects; it has no counterpart in H-theory.

4. The prototype account

In (Horsten, 2019a), an an alternative explication of *ante rem* structuralism in terms of arbitrary objects is given, which we will call the **prototype account** (for reasons that will become clear). Again, we illustrate his approach on the basis of what it says about the natural number structure.

The prototype account is simpler than Fine's account. It identifies the natural number structure $\mathbb{N}_H$ with an arbitrary object that can, for any ω-sequence $\underline{N}$, be in the state of being $\underline{N}$, and can be in no other states. Such an arbitrary object is called a *generic* ω-sequence.

Also for the prototype account, a Benacerrafian worry rears its head. In (Horsten, 2019b, section 6.4), it is argued that there are in fact *many* arbitrary ω-sequences. It is then not clear what is meant when arithmetic is said to have *the* natural number structure as its subject matter. We will return to this problem in section 6.

[11] See (Horsten, 2019b, section 7.5, and p. 158), (Horsten, 2019a, p. 373).
[12] On this point, Fine's view has evolved since Fine (1985).

To this account of the natural number structure, the prototype account adds an account of the individual natural numbers. The natural number n, in this account, is identified with the arbitrary object that, in the state where $\mathbb{N}_H$ takes the value $\underline{N}$, takes the value of the n-th element in this ω-sequence.

This looks much like Fine's account of the individual natural numbers. Indeed, when we compare it with Fine's account of the natural number structure, then, if we disregard Fine's qualification of 'independence' (which is not a primitive notion in H-theory), we can say that the prototype account simply identifies the natural number structure with what Fine calls the *prototype* N that we have encountered in our description of Stage I of Fine's account of the natural numbers structure. Moreover, the account of the individual natural numbers is no more than a copy of Fine's stage II. But the prototype account does *not* go on to build an ω-sequence out of the individual natural numbers and identify the natural number structure with that ω-sequence. In other words, Fine's stage III is dropped altogether.

We immediately see that:

Proposition 4.1. $\mathbb{N}_H$ *is not an* ω*-sequence.*

So, in contrast to Fine's theory of the natural number structure, the prototype account of natural numbers is *non-representational*, whereby a Hellman-style permutation objection cannot be formulated against it.

Observe, incidentally, that a version of the prototype account can be held within Fine's metaphysical framework also. Indeed, I submit that Fine might be well-advised to do so if he does not want to be open to the permutation problem. Of course, this means abandoning the ambition of giving a *representational* account of the natural numbers.

5. Non-rigid structures

The natural number structure is rigid: it allows no non-trivial automorphisms. But we know since (Burgess, 1999)—the third book review that plays an important role in this article—that the mathematical structuralist had better not forget about non-rigid structures. Indeed, we will now see that it is not completely straightforward how F-theory and H-theory apply to non-rigid structures. We investigate this question on the basis of a simple example: the complete two-element graph. But as with the natural number structure, the lessons drawn are intended to generalise.

Consider Fine's account first. We start, in Stage I, by considering the prototype G_2, which is the independent arbitrary object that takes every complete two-element graph *system* as one of its values (and takes no other values). Ultimately, Fine wants G_2 *itself* to be a complete 2-element graph. So what are its two elements a, b? In analogy with the individual natural numbers, we want a^{G_2} to be a dependent arbitrary object which, for any value $\underline{G}$ of G_2, takes one of its two elements as its value. But *which element* of $\underline{G}$ takes a as its value? Any choice here seems completely arbitrary. Of course, if it is not clear what a, b are, then it is also not clear what the complete two-element graph is, on Fine's account.

On the non-representational prototype account, there is no mystery about what the complete two-element graph is: it is (roughly) what Fine calls the prototype G_2. Just as on the prototype account the natural number structure is not itself an ω-sequence, G_2 is not itself a two-element graph. Nonetheless, the prototype account faces a similar problem to the one that Fine faces.

One would like to say that even though it is not a two-element graph in the set theoretic sense of the word, G_2 somehow contains exactly two elements a, b, which are themselves arbitrary objects. But it is not immediately clear what these elements are. Consider a *system* S_1 that instantiates the complete two-element graph. S_1 then consists of two elements a_1, b_1 that stand in a (total) relation to each other. It seems natural to take a, b to be arbitrary objects that can be in a state such that a is a_1, and b is b_1. But why does a take the value a_1 rather than the value b_1? Again, there seems no good answer to this question.

Fine addresses the problem with non-rigid structures in (Fine, 2022). He proposes to extend his account so that arbitrary objects are also allowed to be *multi-valued*. Given this amendation, concerning the complete two-element graph, Fine describes the nature of its two elements v and w as follows (Fine, 2022, p. 613):

> [...] in the particular complete two-element graph G_0 with vertices v_0 and w_0 as a value for [the prototype] G, v will take v_0 and w take w_0 as a value but also, v will take w_0 and w take v_0 as a value.

Fine rightly observes (Fine, 2022, p. 613) that the prototype account can also make use of this solution.[13] The proponent of the prototype theory then simply says that, in the example under consideration, there are *two* situations in which the relevant prototype is in the state of being G_0. One of them is such that v is in the state of being v_0 and w is in the state of being w_0, and the other is such that v is in the state of being w_0 and w is in the state of being v_0. That seems a perfectly satisfactory solution of the problem. Let us therefore amend the prototype theory by adopting Fine's recommendation.

6. Uniqueness

We saw earlier that according to (Horsten, 2019b) and (Horsten, 2019a), there is not just one, but rather there are many arbitrary ω-sequences; moreover, there is no priviliged 'initial' one. On this account, arbitrary ω-sequences together form a *structure* of entities that share a state space. For any arbitrary ω-sequence A_1 in this structure, there is another ω-sequence A_1 such that, in some state s, A_1 and A_2 'are' different particular ω-sequences.

If that is so, then it is not clear how arithmetic can be said to be about a unique structure. Concerning this, I wrote that it is, in a way, a matter of perspective (Horsten, 2019b, p. 112):

> The apparent conflict results from a difference between regarding **N** as a structure or universe on the one hand, and regarding **N** as an element of a larger structure or universe on the other hand. [...] [I]f you are doing number theory (without making use of 'higher mathematics' to obtain number theoretic results), then you are working within *the* generic natural number structure. But the generic natural number structure is itself an entity belonging to a larger universe. If you are making use of generic ω-sequences or are investigating them as a class, then you are dealing with multiple copies of the generic natural number structure that are modally connected with each other.

[13] In (Horsten, 2019a, section 6.8), an alternative solution is proposed for the problem that non-rigid structures pose for an arbitrary objects-account of mathematical structures. This solution proposal was rightly dismissed by Fine as unsatisfactory (Fine, 2022, p. 612–613), and will not be discussed here.

Fine demurs (Fine, 2022, p. 608):

> But the *ante rem* structuralist (as opposed to his eliminative counterpart) is not someone who thinks that it is only relative to a certain perspective that we might talk of the structure of natural numbers as one, even though in fact it is many; and nor, it seems to me [i.e., Fine], is this the attitude of the mathematician towards the complete two-element graph. Indeed, if this were their view, then it is not at all clear why they should not have adopted the Berkeleyan [i.e., *in rebus*] line right from the start and talked about a particular ω-sequence or a particular complete two-element graph as if it were the exemplar of all ω-sequences or all complete two-element graphs.

Certainly Fine has a point. Observe also that Fine's objection affects the prototype account of non-rigid structures as much as its account of rigid structures. Therefore, the prototype accounts at least needs to be further amended. I will now seek to do so.

There is in H-theory a metaphysical correlate to what Fine in this quote refers to as the "perspective from which we might talk of the structure of the natural numbers as one".[14] Recall that a *state* (in my use of the term) is such that one or more arbitrary entities can be in it simultaneously. Consider the generic ω-sequence Ω that can be in any (and only such) state s such that in s, Ω "is" some particular ω-sequence *and no other arbitrary object* (in particular, no other generic ω-sequence) *is in that state* s. By the comprehension principle discussed in section 2, this generic ω-sequence Ω exists. Moreover, by the extensionality principle discussed in that same section, Ω is *unique*.

The generic ω-sequence Ω is an example of an *incomplete arbitrary object space* (Horsten, 2019b, p. 55), i.e., an arbitrary object space containing fewer arbitrary objects than states. Incomplete arbitrary object spaces hitherto received almost no attention in H-theory. This may be the reason why no use was made of them in the account of mathematical structures in (Horsten, 2019a).

Ω forms a 1-element system of arbitrary objects that is completely isolated from—i.e., not correlated with—any other arbitrary object.[15] It is thus "independent" from all other generic ω-sequences. In this respect, Ω differs from Fine's initial ω-sequence. No ω-sequences depend on Ω,—except Ω itself, of course—nor do any other arbitrary objects, whereas many arbitrary objects depend on Fine's initial ω-sequence.

Arithmetic is a self-standing mathematical discipline, in the sense that it is about a single, independent structure, which is instantiated by each ω-sequence. The generic sequence Ω is the *unique* arbitrary ω-sequence that does not belong to a $\geq$2-element system of correlated arbitrary ω-sequences. So we are not faced with a version of Benacerraf's problem of arbitrarily having to choose from a collection of equally suitable arbitrary ω-sequences. In this way, the Ur-generic ω-sequence Ω fits the bill perfectly. I propose that we take it to be the subject matter of arithmetic.

We have seen how the arbtirary ω-sequence Ω *induces*, in the Finean sense,[16] a particular ω-sequence Ω^*. It will by now not come as a surprise that I object to identifying Ω^* with the natural number structure, for this would give rise to a version of the permutation problem.

[14] As, if Fine is right that his and my account are mathematically equivalent, one would expect that there would be.
[15] For a formal discussion of how arbitrary objects are organised in systems in H-theory, see (Steinkrauss and Horsten, 2025).
[16] Cfr supra, section 3.

There is also a lesson here for Fine's friendly suggestion, discussed in section 5, for a satisfactory prototype account of non-rigid structures. Consider again the case of the complete two-element graph, but now suppose that there are *two* particular systems

$$S_1 = \langle \{a_1, b_1\}, R_1 \rangle$$
$$S_2 = \langle \{a_2, b_2\}, R_2 \rangle$$

to be considered. Following Fine's suggestion for defining the elements a, b of the prototype complete two-element graph S (and using his terminology of arbitrary objects x taking a value $v(x)$ in a state), we get a state space $\{s_1, s_2, s_3, s_4\}$ such that:

- in $s_1 : v(S) = S_1, v(a) = a_1, v(b) = b_1$;
- in $s_2 : v(S) = S_1, v(a) = b_1, v(b) = a_1$;
- in $s_3 : v(S) = S_2, v(a) = a_2, v(b) = b_2$;
- in $s_4 : v(S) = S_2, v(a) = b_2, v(b) = a_2$.

So far, so good. But someone might object that there are also arbitrary objects a^*, b^* that behave like a, b, respectively, in s_1 and s_2, but "switch roles" in s_3 and s_4. We have $a \neq a^*$ and $b \neq b^*$. But a^*, b^* seem equally good candidates for being the elements of the complete two-element graph. The proper response to this is that it should be part of the *specification* of each s_i that *no arbitrary objects other than* a, b take values in it.

7. On balance

It is now time to take stock. We have been concerned with the application of arbitrary object theory to the problem of the nature of mathematical structures. In particular, we have discussed the relative merits and demerits of Fine's account of mathematical structures on the one hand, and those of the prototype theory of mathematical structures on the other hand.

We have seen how the original version of Fine's account as well as the original version of the prototype account fail to give a satisfactory account of mathematical structures that admit non-trivial automorphisms: both of them contain unmotivated choice points. In a slight amendment of his original view, Fine proposes a satisfactory way out of this problem. Moreover, he rightly states that this solution can and should also be adopted by the prototype theory.

As Fine in addition points out, the prototype theory of mathematical structures is in addition beset by a uniqueness problem. According to the prototype theory in its first incarnation, there are always *many* prototypes, all of which are equally serviceable as *ante rem* mathematical structures. Any choice between them would be arbitrary, and any appeal to an implicit quantification over all of them would push us in the direction of *in rebus* structuralism. In short, we need a notion of prototype that ensures the relevant uniqueness.

In the form of incomplete arbitrary object spaces, H-theory contains the resources to identify the right notion of prototype: they are in some sense analogues in H-theory of Finean *initial* arbitrary objects. The resulting (further) **amended prototype theory** is then no longer vulnerable to Fine's non-uniqueness objection. Indeed, the amended prototype theory seems, at least so far, a perfectly satisfactory arbitrary object theoretical account of mathematical structures.

Because of its representationality, Fine's own amended account of mathematical structures is in addition marred by a version of Hellman's permutation problem. The prototype theory is not

representational and is therefore not vulnerable to any permutation challenges. Moreover, the prototype theory, even in its amended version, is perfectly compatible with F-theory. I therefore recommend Fine to simply abandon representationality and adopt the amended prototype theory also.

At least on the face of it, none of all this tells against the background F-theory or against the background H-theory of arbitrary objects. This is in itself an interesting finding. True, because it has fewer primitive notions, H-theory is more economical. But there may well be decisive reasons for taking the relation of dependence between arbitrary objects as a primitive element of arbitrary object theory. Maybe the future will tell.

Acknowledgements. I am grateful to two anonymous referees for helpful comments, questions, and suggestions for improvement.

References

Burgess, J. (1999). Identity, indiscriminability, and ante rem structuralism. Book Review: Stewart Shapiro, Philosophy of mathematics: structure and ontology. *Notre Dame Journal of Formal Logic*, 40(1):283–291. DOI: https://doi.org/10.1305/ndjfl/1038949543.

Burgess, J. (2008). Critical study / book review: Charles Parsons. Mathematical thought and its objects. *Philosophia Mathematica*, 16:402–420. DOI: https://doi.org/10.1093/philmat/nkn017.

Fine, K. (1985). *Reasoning with arbitrary objects*. Blackwell.

Fine, K. (1998). Cantorian abstraction: a reconstruction and defense. *Journal of Philosophy*, 95:599–634. DOI: https://doi.org/10.5840/jphil1998951230.

Fine, K. (2022). The Metaphysics and Mathematics of Arbitrary Objects, by Leon Horsten. Cambridge: Cambridge University Press, 2019. pp. xviii + 232. *Mind*, 131:603–618. DOI: https://doi.org/10.1093/mind/fzaa059.

Hellman, G. (2006). Structuralism. In Shapiro, S., editor, *The Oxford Handbook of Philosophy of Mathematics and Logic*, pages 536–562. Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780195325928.003.0017.

Horsten, L. (2019a). Generic Structures. *Philosophia Mathematica*, 27:362–380. DOI: https://doi.org/10.1093/philmat/nky015.

Horsten, L. (2019b). *The Metaphysics and Mathematics of Arbitrary Objects*. Cambridge University Press. DOI: https://doi.org/10.1017/9781139600293.

Shapiro, S. (1997). *Philosophy of Mathematics: Structure and Ontology*. Oxford University Press. DOI: https://doi.org/10.1093/0195139305.001.0001.

Steinkrauss, L. and Horsten, L. (2025). Axioms for Arbitrary Object Theory. arXiv. arXiv:2504.20122. DOI: https://doi.org/10.48550/arXiv.2504.20122.

Journal for the Philosophy of Mathematics

Firenze University Press
www.fupress.com

On Class Hierarchies

LUCA INCURVATI

Department of Philosophy and Institute for Logic, Language and Computation, University of Amsterdam.
Email: l.incurvati@uva.nl

OPEN ACCESS

Citation: INCURVATI, Luca (2025). On Class Hierarchies. *Journal for the Philosophy of Mathematics.* 2: 45-74. doi: 10.36253/jpm-3459

Received: April 15, 2025

Accepted: June 22, 2025

Published: December 30, 2025

ORCiD
LI: 0000-0001-7381-7378

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Abstract: In her seminal article 'Proper Classes', Penelope Maddy introduced a novel theory of classes validating the naïve comprehension rules. The theory is based on a step-by-step construction of the extension and anti-extension of the membership predicate, which mirrors Kripke's construction of the extension and anti-extension of the truth predicate. Maddy's theory has been criticized by Øystein Linnebo for its 'rampant indeterminacy' and for making identity among classes too fine-grained. In this paper, I present a theory of classes which, while building on Maddy's theory, avoids its rampant indeterminacy and allows for identity among classes to be suitably coarse-grained. I begin by presenting a bilateral natural deduction system for Maddy's theory, which improves on her axiomatization in several respects. I then go on to show how to avoid the rampant indeterminacy by using supervaluational schemes in the construction of the extension and anti-extension of the membership predicate and how to augment the proof theory with corresponding, motivated rules. It turns out that whilst a van Fraassen-style supervaluational scheme suffices to avoid the basic problem of rampant indeterminacy, a supervaluational scheme based on maximally consistent extensions is needed for a proper treatment of identity.

Keywords: Classes, hierarchies, supervaluation, Maddy, identity, maximal consistency, bilateral logic.

1. Introduction

Despite early doubts about their legitimacy—doubts that, in the philosophical literature, continue to this day—sets have become the bread and butter of mathematics. And despite doubts about its cogency and limitations, the iterative conception of set—according to which sets are the objects obtained by iterated applications of the *set of* operation—has become the received view of sets in mathematics and philosophy.[1] Even if one accepts this received view, one might well wonder: are there collections other than sets?

[1] For an extended defence of the iterative conception against rival conceptions of set, see Incurvati, 2020.

George Boolos forcefully advocated a negative answer—a 'settist' position, as he liked to call it. In response to Charles Parsons's (1974) proposal for a theory of sets *and* classes, he famously complained:

> Wait a minute! I thought that set theory was supposed to be a theory about all, 'absolutely' all, the collections that there were and that 'set' was synonymous with 'collection'. (Boolos, 1974, 35)

Those that defend a positive answer—those who argue that we should countenance collections other than sets—typically do so on the grounds that they are needed to perform tasks that sets are unable or less suited to accomplish. Indeed, it is in part because of his settist convictions that Boolos embarked upon his programme of providing a plural interpretation of second-order logic: our apparent talk about what appear to be collections that cannot form a set, such as the collection of all ordinals and the collection of all sets, should be seen as plural talk in disguise.

One of the philosophers to have most clearly presented some of the reasons for admitting collections other than sets is Penelope Maddy. In her seminal article 'Proper classes' (Maddy, 1983), she marshalled a series of mathematical and philosophical considerations in support of the existence of classes and indeed proper classes—collections that are too big to form a set. On the basis of these considerations, Maddy went on to introduce a novel theory of classes, whose mathematical properties she investigated further in later work (Maddy, 2000).[2]

Maddy's theory is based on a step-by-step construction of the extension and anti-extension of the membership predicate, which mirrors Kripke's (1975) construction of the extension and anti-extension of the truth predicate. The theory has been criticized by Øystein Linnebo for its 'rampant indeterminacy' and for making identity among classes too fine-grained. In this paper, I present a theory of classes that builds on Maddy's theory but avoids its rampant indeterminacy and allows for identity among classes to be suitably coarse-grained.

I begin by providing an overview of the roles that classes have been invoked to play and isolating some key desiderata on a theory of classes to which these roles give rise. I then present Maddy's theory of classes and provide a bilateral natural deduction system for the theory, which improves on Maddy's own axiomatization in several respects. I go on to describe the problem of rampant indeterminacy for Maddy's theory, and I show how it can be avoided by using supervaluational schemes in the construction of the extension and anti-extension of the membership predicate. It turns out, however, that whilst a van Fraassen-style supervaluational scheme suffices to avoid the basic problem of rampant indeterminacy, a supervaluational scheme based on maximally consistent extensions is needed for a proper treatment of identity. For both supevaluational theories, I also provide bilateral natural deduction systems. Indeed, the proof theory will play a central role in motivating the treatment of identity and in answering a challenge of Greg Restall (2010) to any coarse-grained account of identity in naïve class theory.

2. Desiderata on classes

Classes have been invoked to play a variety of roles. A detailed examination and assessment of these roles is beyond the scope of this paper. My goal is to provide an overview of the most

[2] As Maddy (2024) points out, the publication date of Maddy, 2000 is misleading: it was written in the mid-80s. It should also be noted that Maddy has since then come closer to endorsing a settist position herself, on the grounds that classes do not appear to provide any real mathematical contribution to the theory of sets. Nonetheless, she is open to the possibility of classes being needed for different endeavours, such as formal semantics. See Maddy, 2024, 82–83.

prominent roles, with the aim of lending support to the desiderata on a theory of classes we will rely on.[3]

In general, classes have been called upon to play roles that sets have been deemed unsuited to serve. First, it has been argued that classes are *genuinely used* in set-theoretic practice (Barton and Williams, 2024; Linnebo, 2006; Parsons, 1974). Upon opening a set theory textbook, one immediately comes across terms such as V, L and Ω, which intuitively stand for collections (namely, the collection of all sets, the collection of all constructible sets and the collection of all ordinals) that are too big to form a set and should therefore be considered classes. It is however common place to regard this apparent talk of classes as a a way of stating facts that only involve sets. For instance, when a set theorist writes '$x \in \Omega$', this ought to be taken as a shorthand for 'x is a transitive set linearly ordered by $\in$'. To use Quinean terminology (Quine, 1948), classes are convenient myths, to be paraphrased away. However, it is not clear how far the paraphrase strategy can be taken. A case that is especially difficult to handle concerns the formulation and treatment of certain large cardinal axioms and reflection principles. For instance, the paraphrase strategy applied to non-trivial elementary embeddings would seem to trivialize Kunen's celebrate theorem that there is no non-trivial elementary embedding of the universe onto itself (Fujimoto, 2019; Hamkins et al., 2012).[4]

Second, it has been argued that classes are needed to *make sense of* set-theoretic practice. A case is point is the debate on unrestricted quantification (Florio, 2014). In standard model theory, an interpretation is an ordered pair consisting of the domain and an interpretation function, where both of these are sets. But according to the iterative conception, there is no set of all sets. Hence, there can be no interpretation whose domain contains all sets. But this is problematic. For one thing, if the intended interpretation of set theory is the one in which the domain contains all sets, and if truth is truth on the intended interpretation, then it would seem impossible to account for the apparent truth of 'Every set has a power set' or 'No set contains all sets'. For another, if we define logical truth as truth on all interpretations, then a sentence whose quantifiers range over all sets could be logically true and yet false in the universe of sets (Kreisel, 1967).

The conclusion that some (e.g., Lear, 1977; Parsons, 1974) have been willing to draw is that the set-theoretic quantifiers never range over absolutely all sets, and so there is not a single, intended interpretation of set theory. Others have tried to resist this conclusion by jettisoning the assumption that the domain of quantification is a set. But if not a set, what is it? One option is that the domain is not a single entity at all: it is a plurality of things—*the sets* (Boolos, 1985; Rayo and Uzquiano, 1999). Another option is that it is not an object but a higher-order entity of some kind—perhaps a Fregean concept (Rayo and Williamson, 2003). Yet another option is that the domain is a collection-like object other than a set—a class. Classes might therefore allow us to preserve both the idea that the set-theoretic quantifiers range over all sets and that the domain of quantification is an object, as standard model-theoretic semantics would have it.

Third, it has been argued that classes are needed for a treatment of cardinality in line with our basic intuitions about numbers (Maddy, 1983). Cardinal numbers are numbers that represent the size of collections. According to Cantor, a theory of size should be based on the concept of one-to-one correspondence: the cardinality of two sets should be the same just in case they can be put into one-to-one correspondence—a version of Hume's Principle (Frege, 1884). It is

[3] See also Schindler, 2019 for discussion of the roles that classes have been invoked to play.

[4] Another difficult case concerns work on class forcing. See Gitman et al., 2020; Holy et al., 2016.

then natural to take the cardinality of a set a to *just be* the collection of sets that can be put into one-to-one correspondence with a. This definition immediately delivers Hume's Principle, and it is known as the Frege–Russell definition of cardinal number (Frege, 1884; Russell, 1903; see Incurvati, 2020, 33 for discussion). However, the collection of sets that can be put into one-to-one correspondence with a given set is not, in general, a set. It could be a class, however. Classes, therefore, provide a way of rehabilitating a Frege–Russell treatment of cardinality.[5] Similar considerations apply, *mutatis mutandis*, to the development of ordinal numbers by replacing the notion of one-to-one correspondence with the notion of an isomorphism.

Fourth, it has been argued that classes are needed for the development of natural language semantics. In standard model-theoretic semantics, the predicate 'is wise' in 'Andrea is wise' is standardly taken to pick out a function representing a property. However, in English we also have nominalizations of predicates (such as 'wisdom' in 'Wisdom is a virtue'), which would seem to purport to refer to properties in nominal position. It is widely acknowledged that nominalizations are very useful, and there are strong reasons for thinking they are not always eliminable (see Button and Trueman, 2024). One may be tempted to take this to show that properties are objects, and theories that treat properties as objects have been devised (Aczel, 1980; Bealer, 1982). However, there is a longstanding philosophical tradition, going back to Frege (1892) and recently revived by Jones (2016) and Trueman (2021), for regarding the idea of treating properties as objects as incoherent. But even if properties are not objects, there seems still to be the need for objects which serve as first-order proxies for properties and be the referents of nominalizations (Chierchia and Turner, 1988, Incurvati, 2020: Ch. 7). These objects cannot be sets since, among the predicates of our language, there are predicates such as 'is a set' or 'is an ordinal', whose nominalizations cannot refer to sets. Thus, the argument concludes, the referents of nominalizations must be classes.

Having described some key roles that classes have been invoked to play, we can now list a numbers of basic desiderata sanctioned by these roles. First, our discussion should make it clear that there should be 'big' classes such as the class of all ordinals or the class of all sets. Ideally, we would want to validate the naïve comprehension principle that to every condition there corresponds a class of all and only the things satisfying that condition.

Second, classes should be *logical* collections: they are characterized by reference to some predicate, concept or property, which determines its members. This is in contrast with sets, which are *combinatorial* collections: they are characterized not by reference to some predicate, concept, or property, but by reference to their members in a more direct fashion.[6] For one thing, as our discussion has made clear (especially the discussion of the fourth role that classes have been invoked to play), there seems to be a tight connection between classes and predicates or properties, which ought to be vindicated by what we take classes to be. For another, as Maddy (1983, 122) emphasizes, it should be clear why classes are not just 'another stage of sets we forgot to include'. Providing a theory of classes as logical collections allows us to clearly distinguish them from iterative sets, which are combinatorial collections.

Third, as Maddy (1983, 120–123) also demands, classes should be real, well-defined entities. Our discussion of the first role of classes makes it clear why, metaphysical considerations aside,

[5] This is not to say that other strategies are not available. In particular, it is still possible to approximate the Frege–Russell definition within iterative set theory by considering the collection of sets that can be put into one-to-one correspondence with a that occur as low as possible in the cumulative hierarchy. This is known as the Scott-Tarski definition of a cardinal. See Incurvati, 2020, 80.

[6] On the distinction between logical and combinatorial collections, see Maddy, 1990, 121 and Incurvati, 2020, 31.

this is a sensible desideratum on a theory of classes. For if classes are not real entities to be eliminated via paraphrase, it becomes hard to make sense of some aspects of set-theoretic practice such as Kunen's theorem about elementary embeddings.

Finally, if classes are real objects, it should be possible to provide a satisfactory criterion of identity for classes. No entity without identity, after all. Now, *qua* logical collections, it may be unreasonable or even wrong to demand that classes ought to be extensional: identity among classes ought not to be too coarse-grained. But identity among classes ought not be too fine-grained either. I will not belabour this point further, because we will return to it in greater detail below.

3. Maddy's theory of classes

We can now describe Maddy's (1983; 2000) theory of classes. Her key idea is to address the logical paradoxes by allowing certain membership claims involving classes to be indeterminate. Kripke's (1975) theory of truth famously allows certain truth claims to be indeterminate. He outlines his theory by presenting a construction in which the extension and the anti-extension of the truth predicate (that is, the collection of things to which the predicate definitely applies and the collection of things to which the predicate definitely does not apply) are built in stages. Maddy presents her theory of classes by specifying a construction in which the extension and the anti-extension of the membership relation are similarly built in stages. To emphasize the analogies with Kripke's construction, which will be relevant below, I will follow the presentation of Maddy's theory recently provided by Linnebo (2024).

We begin by specifying the object language of the theory. Ultimately, we are going to want to use membership relations involving only sets, which we take as given, to determine membership relations involving classes as well as sets. One option would be to think of these as two different relations; another option is to treat them as the same relation holding between different kinds of entities. Maddy opts for the second option.[7] Thus, we start with the language of first-order logic with identity extended with a membership symbol η (reserving the symbol $\in$ for membership in the meta-language).

Next, we want to have terms for the objects of our theory. We have a hat operator which produces a class term when applied to an open formula. So we have class terms such as $\hat{x}(x = x)$ and $\hat{x}(x \not\eta x)$. To increase the expressive power of the language, Maddy adds a class constant $\bar{V}$, standing for the universe of sets V. On Maddy's account, V is simply the real universe of sets, so it will be a class in the meta-theory. For reasons that will become clear below, it is in fact preferable to let $\bar{V}$ denote a set from the point of view of the meta-theory, which serves as the universe of sets and hence a class from the point of view of the object theory. A natural choice, and the one which we shall stick to, is to let $\bar{V}$ denote V_κ, where κ is the first inaccessible cardinal. Accordingly, our meta-theory will consist of ZFC plus the assertion that there exists an inaccessible cardinal. Finally, Maddy assumes that for each $a \in V$ there is a constant $\bar{a}$.

Formally, the terms and formulae of the language $\mathcal{L}_\eta$ are defined as follows:

(i) All constants and variables are terms.
(ii) If t and t' are terms, then $t = t'$ and $t \eta t'$ are formulae.
(iii) If F and G are formulae and x is a variable, then $(F \wedge G)$, $\neg F$ and $\forall x\, F$ are formulae.

[7] As Linnebo (2024, 72) notes, not much hinges on the choice from a technical point of view, and it would be possible to reformulate (with some adjustments) the discussion to follow using the first option.

(iv) If F is a formula and x is among the free variables of F, then $\hat{x}F$ is a term.

Sentences are closed formulae. In general, F and G stand for (possibly open) formulae, whereas A and B always stand for sentences. We let T be the set of all terms of $\mathcal{L}_\eta$, T^* be the set of all closed terms (a subset of T), C be the set of all class terms and C^* be the set of all closed class terms (a subset of C). Disjunction, the conditional, the biconditional, and the existential quantifier are defined as usual. That is, $F \vee G$ is defined as $\neg(\neg F \wedge \neg G)$, $F \rightarrow G$ is defined as $\neg(F \wedge \neg G)$, $F \leftrightarrow G$ is defined as $(F \rightarrow G) \wedge (G \rightarrow F)$, and $\exists x F$ is defined as $\neg\forall x \neg F$.

Our next step is to provide a model theory for the language. We represent an extension σ^+ and anti-extension σ^- of η as sets of ordered pairs of closed terms, that is subsets of $T^* \times T^*$. Any pair σ of such an extension and anti-extension can be regarded as a model of the Strong Kleene Logic K_3 provided that $\sigma^+ \cap \sigma^- = \emptyset$. In particular, we can recursively define a satisfaction relation $\models$ indicating what the model thinks is true of sets and classes, and an anti-satisfaction relation $=\!\!|$ indicating what the model thinks is false of them:

- $\sigma \models t = t'$ iff $t = t'$ for all $t, t' \in T^*$;
- $\sigma =\!\!| \, t = t'$ iff $t \neq t'$ for all $t, t' \in T^*$;
- $\sigma \models t\eta t'$ iff $\langle t, t' \rangle \in \sigma^+$ for all $t, t' \in T^*$;
- $\sigma =\!\!| \, t\eta t'$ iff $\langle t, t' \rangle \in \sigma^-$ for all $t, t' \in T^*$;
- $\sigma \models \neg A$ iff $\sigma =\!\!| \, A$;
- $\sigma =\!\!| \, \neg A$ iff $\sigma \models A$;
- $\sigma \models A \wedge B$ iff $\sigma \models A$ and $\sigma \models B$;
- $\sigma =\!\!| \, A \wedge B$ iff $\sigma =\!\!| \, A$ or $\sigma =\!\!| \, B$;
- $\sigma \models \forall x \, F$ iff $\sigma \models F[t/x]$ for all $t \in T^*$.
- $\sigma =\!\!| \, \forall x \, F$ iff $\sigma =\!\!| \, F[t/x]$ for some $t \in T^*$.

It is straightforward to show that, similarly to the case of Kripke's theory of truth, satisfaction and anti-satisfaction are monotonic in the following sense: if σ_2 is an *expansion* of σ_1 (in symbols: $\sigma_1 \sqsubseteq \sigma_2$, where $\sigma_1 \sqsubseteq \sigma_2$ iff $\sigma_1^+ \subseteq \sigma_2^+$ and $\sigma_1^- \subseteq \sigma_2^-$), then, for any A, $\sigma_1 \models A$ only if $\sigma_2 \models A$, and $\sigma_1 =\!\!| \, A$ only if $\sigma_2 =\!\!| \, A$.

The last step is that of building a sufficiently encompassing extension/anti-extension pair (EA-pair for short). At stage zero, we have all positive and negative membership facts given from set theory and the intuitive meaning of $\bar{V}$. At limit stages, we take unions as customary. At successor stages, just as in the case of Kripke's construction, we apply a jump operation J taking us from an EA-pair σ to another in a monotonic fashion, i.e. so that $J(\sigma)$ is an expansion of σ. Maddy opts for a jump operation which is the exact analogue of Kripke's jump operation for truth: if a term t satisfies a formula F, we add $\langle t, \hat{x}F \rangle$ to the extension of η; if t anti-satisfies F, we add it to the anti-extension.

Definition 3.1 (Maddy jump)**.** Given an EA-pair $\sigma = \langle \sigma^+, \sigma^- \rangle$, the *Maddy jump* J_{M} is the operation such that, for all $t \in T^*$, $J_{\mathrm{M}}(\sigma^+) = \{\langle t, \hat{x}F \rangle \mid \sigma \models Ft\}$ and $J_{\mathrm{M}}(\sigma^-) = \{\langle t, \hat{x}F \rangle \mid \sigma =\!\!| \, Ft\}$.

We are now ready to recursively define the *Maddy hierarchy*.

Definition 3.2 (The Maddy hierarchy)**.**

$$\sigma_0^{+,\mathrm{M}} = \{\langle \bar{a}, \bar{b}\rangle \mid V_\kappa \models a \in b\} \cup \{\langle \bar{a}, \bar{V}\rangle \mid a \in V_\kappa\};$$
$$\sigma_0^{-,\mathrm{M}} = \{\langle \bar{a}, \bar{b}\rangle \mid V_\kappa \models a \notin b\} \cup \{\langle t, \bar{a}\rangle \mid a \in V_\kappa\} \cup \{\langle t, \bar{V}\rangle\} \text{ for all } t \in C^*;$$
$$\sigma_{\alpha+1}^{+,\mathrm{M}} = J^+{}_{\mathrm{M}}(\sigma_\alpha);$$
$$\sigma_{\alpha+1}^{-,\mathrm{M}} = J^-{}_{\mathrm{M}}(\sigma_\alpha);$$
$$\sigma_\lambda^{+,\mathrm{M}} = \bigcup_{\alpha<\lambda} \sigma_\alpha^{+,\mathrm{M}} \text{ if } \lambda \text{ is a limit ordinal;}$$
$$\sigma_\lambda^{-,\mathrm{M}} = \bigcup_{\alpha<\lambda} \sigma_\alpha^{-,\mathrm{M}} \text{ if } \lambda \text{ is a limit ordinal.}$$

By a theorem of Flagg (who established a conjecture of Tait's, see Maddy, 2000, 315, fn. 22), the construction of the Maddy hierarchy where V is a set in the meta-theory reaches a fixed point σ^{M} at the first admissible ordinal greater than all the ordinals in V.[8] We can therefore consider σ^{M} as giving the extension and anti-extension of η according to Maddy's theory of classes.

4. A bilateral system for Maddy's theory

At the end of her 1983 paper, Maddy posed the following question:

Question 4.1 (Maddy)**.** How and to what extent can the theory of V^* [the universe of sets and classes in Maddy's theory] be axiomatized?

I am going to provide a sound and complete axiomatization of the fixed point σ^{M} of the Maddy hierarchy. The axiomatization builds on the axiomatization provided by Maddy (2000) (with suggestions from Myhill), but also differs from it in important respects. First, its background logic is bilateral, in that it uses signs for the speech acts of assertion and rejection. This will allow me to formulate the logic in a natural deduction system with rules that are harmonious and separable. Second, the axiomatization allows one to *reason* using Maddy's theory of classes, whereas Maddy's axiomatization only allows one to compute what is in σ^{M}. Concretely, this will put us in a position to provide a general model theory for which the natural deduction system is sound and complete. Finally, the natural deduction system will pave the way towards the theories of classes that I will develop below to address the problems with Maddy's theory.

We begin by characterizing the language of the system. As announced, we are working towards a bilateral system. So we extend the language $\mathcal{L}_\eta$ to the signed language $\mathcal{L}_\eta^S$ by including the signs $+$ and $-$, standing, respectively, for the speech acts of assertion and rejection. The notions of term, formula and sentence are defined as in $\mathcal{L}_\eta$, but, in addition, we also have the notion of a signed sentence, which will be anything obtained by prefixing a sentence with a $+$ or a $-$.

We can now lay down the rules of the natural deduction system. The model-theoretic clauses for conjunction tell us that a conjunction is satisfied just in case both of its conjuncts are, and anti-satisfied just in case at least one of its conjunct is. The proof-theoretic side of this coin,

[8] By contrast, as Maddy (2000, 308) notices and was proved by Tait, the construction has no fixed point if V is the universe of sets from the point of view of the meta-theory.

in terms of assertion and rejection, is captured by the following rules (where φ is a signed sentence):

$$(+\wedge\text{I.})\ \frac{+A \quad +B}{+A \wedge B} \qquad (+\wedge\text{E.}_1)\ \frac{+A \wedge B}{+A} \qquad (+\wedge\text{E.}_2)\ \frac{+A \wedge B}{+B}$$

$$(-\wedge\text{I.}_1)\ \frac{-A}{-A \wedge B} \qquad (-\wedge\text{I.}_2)\ \frac{-B}{-A \wedge B} \qquad (-\wedge\text{E.})\ \frac{-A \wedge B \quad \begin{array}{c}[-A]\\ \vdots\\ \varphi\end{array} \quad \begin{array}{c}[-B]\\ \vdots\\ \varphi\end{array}}{\varphi}$$

The model-theoretic clauses for satisfaction and anti-satisfaction of negated and universally quantified sentences can be similarly mirrored by the following natural deduction rules for the assertion and rejection of those sentences:

$$(+\neg\text{I.})\ \frac{-A}{+\neg A} \qquad (+\neg\text{E.})\ \frac{+\neg A}{-A} \qquad (-\neg\text{I.})\ \frac{+A}{-\neg A} \qquad (-\neg\text{E.})\ \frac{-\neg A}{+A}$$

$$(+\forall\text{I.})\ \frac{+F[t/x]}{+\forall x\, F}\ \text{for all } t \in T^* \qquad (+\forall\text{E.})\ \frac{+\forall x\, F}{+F[t/x]}\ \text{for any } t \in T^*$$

$$(-\forall\text{I.})\ \frac{-F[t/x]}{-\forall x\, F}\ \text{for some } t \in T^* \qquad (-\forall\text{E.})\ \frac{-\forall x\, F \quad \begin{array}{c}[-F[t/x]]\\ \vdots\\ \varphi\end{array}}{\varphi}\ \text{for any } t \in T^*$$

Note that the quantifiers rules are simply infinitary counterparts of the conjunction rules. The infinitary character of these rules should not be entirely surprising, given that we are attempting to characterize the fixed point of a hierarchy whose first stage includes all positive and negative membership facts settled by V_κ. Indeed, the situation is analogous to the situation for truth. In that case, the first stage of the relevant hierarchies (such as the Kripke hierarchy and supervaluational versions thereof) includes all facts settled by arithmetic. For this reason, axiomatizations of the fixed point of this hierarchies typically include some version of the ω-rule (see Incurvati and Schlöder, 2023; Meadows, 2015).

We now turn to the treatment of identity. Identity is a thorny issue in the context of naïve theories of classes, and we shall return to it below. For now, recall that Maddy's model-theoretic clauses for identity tell us that $t = t$ is always satisfied and $t = t'$ is always anti-satisfied. These clauses can be mirrored in the proof theory by the following introduction rules.

$$(+=\text{I.})\ \frac{}{+t = t} \qquad (-=\text{I.})\ \frac{}{-t = t'}\ \text{for any distinct } t, t' \in T^*$$

How could we formulate suitable elimination rules for identity? The introduction rules tell us that nothing is required to infer $+t = t$ and $-t = t'$. Accordingly, nothing can be inferred from those statements either. Nonetheless, given that these are *all* the rules for identity we also know that we can never infer $-t = t$ and $+t = t'$. So assuming that the logic of the meta-theory encompasses minimal logic, it follows that anything can be inferred from them.

$$(-=\text{E.})\ \frac{-t=t}{\varphi} \qquad (+=\text{E.})\ \frac{+t=t'}{\varphi}\ \text{for any distinct } t, t' \in T^*$$

Kripke's construction was designed to validate the naïve truth rules allowing us to move between the assertion of A and that of 'A is true'. Similarly, Maddy's construction was designed to validate the naïve class rules allowing us to move between the assertion of Ft and the assertion of 't is in $\hat{x}F$'. Similar considerations apply for the case in which Ft and 't is in $\hat{x}F$' are rejected. We therefore add the following rules:

$$(+\eta\text{I.})\ \frac{+Ft}{+t\eta\hat{x}F} \qquad (+\eta\text{E.})\ \frac{+t\eta\hat{x}F}{+Ft} \qquad (-\eta\text{I.})\ \frac{-Ft}{-t\eta\hat{x}F} \qquad (-\eta\text{E.})\ \frac{-t\eta\hat{x}F}{-Ft}$$

The rules laid down so far (indeed, the introduction rules alone) would suffice to characterize the fixed point of the Maddy hierarchy. Recall, however, that we want to provide a deductive system that allows us to reason using Maddy's theory. Recall, moreover, that any EA-pair can be considered a K_3 model. But K_3 is an explosive logic, in which anything follows from a contradiction. In a bilateral context, this suggests adding the following rule of bilateral explosion, as Ryan Simonelli (*forthcoming*) has suggested:

$$\text{(Bilateral Explosion)}\ \frac{+A \qquad -A}{\varphi}$$

Note that the identity elimination rules are an immediate consequence of the identity introduction rules and Bilateral Explosion. Bilateral Explosion does not serve to specify the meaning of a particular logical constant, but rather to characterize the interaction between the speech acts of assertion and rejection in the proof theory. Principles governing the interaction between speech acts are known as coordination principles in the bilateral and multilateral literature (Incurvati and Schlöder, 2023; Rumfitt, 2000; Smiley, 1996).

Let Maddian Bilateral Logic (MBL for short) be the system consisting of the above introduction and elimination rules for negation, conjunction, the universal quantifier, identity and membership together with the Bilateral Explosion rule. To be able to characterize the fixed point of the Maddy hierarchy, one last ingredient is needed: we need to give a theory providing all the facts about membership among sets and about membership in $\bar{V}$ from which the construction of the Maddy hierarchy begins. So let $V_\kappa \models A$ be defined as usual for As in the language $\mathcal{L}_\in$ of set theory. True Set Theory (TST for short) is then the union of the following sets of signed sentences:

(i) $\{+\bar{a}\eta\bar{b} \mid V_\kappa \models a \in b\}$
(ii) $\{-\bar{a}\eta\bar{b} \mid V_\kappa \models a \notin b\}$
(iii) $\{+\bar{a}\eta\bar{V} \mid a \in V_\kappa\}$
(iv) $\{-t\eta\bar{a} \mid a \in V_\kappa\}$ for all $t \in C^*$
(v) $\{-t\eta\bar{V}\}$ for all $t \in C^*$

We are now in a position to prove that, over TST, MBL axiomatizes the fixed point of the Maddy hierarchy. We begin by showing the soundness of MBL with respect to the model-theoretic consequence relation induced by σ^{M}.

Theorem 4.2. *Let Γ be a set of signed sentences and φ a signed sentence. Suppose that $\Gamma \vdash_{\mathsf{MBL}} \varphi$ and that for all ψ in Γ, $\sigma^{\mathrm{M}} \models A$ if $\psi = +A$, and $\sigma^{\mathrm{M}} \models\!\!| A$ if $\psi = -A$. Then, $\sigma^{\mathrm{M}} \models B$ if $\varphi = +B$ and $\sigma^{\mathrm{M}} \models\!\!| B$ if $\varphi = -B$.*

Proof. The identity introduction rule are sound because for any distinct $t, t' \in T^*$, $\sigma^{\mathrm{M}} \models t = t$ and $\sigma^{\mathrm{M}} \models\!\!| t \neq t'$. Moreover, it is easy to verify that if the premises of the rules for the connectives, the universal quantifier and η are (anti-)satisfied by σ^{M}, then so is the conclusion. (For the η-introduction rules, we use the definition of the Maddy jump and the fact that σ^{M} is a fixed point.) Finally, the Rejection rule is sound because by construction of the Maddy hierarchy, for no A it is ever the case that $\sigma^{\mathrm{M}} \models A$ and $\sigma^{\mathrm{M}} \models\!\!| A$. □

Since by construction $\sigma^{\mathrm{M}} \models A$ for all $+A \in \mathsf{TST}$ and $\sigma^{\mathrm{M}} \models\!\!| A$ for all $-A \in \mathsf{TST}$, Theorem 4.2 immediately yields that, over TST, MBL is sound with respect to σ^{M}.

Theorem 4.3. *For every sentence A, if $\mathsf{TST} \vdash_{\mathsf{MBL}} +A$, then $\sigma^{\mathrm{M}} \models A$, and if $\mathsf{TST} \vdash_{\mathsf{MBL}} -A$, then $\sigma^{\mathrm{M}} \models\!\!| A$.*

I now prove that, over TST, MBL is also complete with respect to σ^{M}.

Theorem 4.4. *For every sentence A, if $\sigma^{\mathrm{M}} \models A$, then $\mathsf{TST} \vdash_{\mathsf{MBL}} +A$, and if $\sigma^{\mathrm{M}} \models\!\!| A$, then $\mathsf{TST} \vdash_{\mathsf{MBL}} -A$.*

Proof. I prove that if $\sigma^{\mathrm{M}}_{\alpha} \models A$, then $\mathsf{TST} \vdash_{\mathsf{MBL}} +A$, and if $\sigma^{\mathrm{M}}_{\alpha} \models\!\!| A$, then $\mathsf{TST} \vdash_{\mathsf{MBL}} -A$. The theorem will then follow by ordinal induction.
Base case. Immediate from the definition of TST.
Inductive step. Our induction hypothesis is that, for every A, if $\sigma^{\mathrm{M}}_{\alpha} \models A$, then $\mathsf{TST} \vdash_{\mathsf{MBL}} +A$, and if $\sigma^{\mathrm{M}}_{\alpha} \models\!\!| A$, then $\mathsf{TST} \vdash_{\mathsf{MBL}} -A$. We need to prove that, for every A, if $\sigma^{\mathrm{M}}_{\alpha+1} \models A$, then $\mathsf{TST} \vdash_{\mathsf{MBL}} +A$, and if $\sigma^{\mathrm{M}}_{\alpha+1} \models\!\!| A$, then $\mathsf{TST} \vdash_{\mathsf{MBL}} -A$. We proceed by induction on the complexity of A. We only cover the positive cases, since the negative cases proceed similarly, using the negative introduction rule for the relevant logical constant, rather than the positive one.

A is of the form $t\eta t'$. Suppose $\sigma^{\mathrm{M}}_{\alpha+1} \models t\eta t'$. If t' is a set term or $\bar{V}$, we have that $\mathsf{TST} \vdash_{\mathsf{MBL}} +t\eta t'$ by the fact that $\sigma_{\alpha+1} \sqsupseteq \sigma_0$. So suppose that t' is of the form $\hat{x}F$ for some F. Then, by construction of the Maddy hierarchy, $\sigma^{\mathrm{M}}_{\alpha} \models Ft$. By the induction hypothesis of the ordinal induction, it follows that $\mathsf{TST} \vdash_{\mathsf{MBL}} +Ft$. But then, $\mathsf{TST} \vdash_{\mathsf{MBL}} +t\eta\hat{x}F$ since MBL includes the ($+\eta$I.) rule.

A is of the form $t = t'$. If $\sigma^{\mathrm{M}}_{\alpha+1} \models t = t'$, then, by the model-theoretic clauses for identity, t is the same term as t'. But then, $\mathsf{TST} \vdash_{\mathsf{MBL}} +t = t'$ since MBL includes the ($+=$I.) rule.

A is of the form $B \wedge C$. If $\sigma^{\mathrm{M}}_{\alpha+1} \models B \wedge C$, then, by the model-theoretic clauses for conjunction, $\sigma^{\mathrm{M}}_{\alpha+1} \models B$ and $\sigma^{\mathrm{M}}_{\alpha+1} \models C$. But then, by the induction hypothesis of the induction on complexity, $\mathsf{TST} \vdash_{\mathsf{MBL}} +B$ and $\mathsf{TST} \vdash_{\mathsf{MBL}} +C$. Since MBL includes the ($+\wedge$I.) rule, it follows that $\mathsf{TST} \vdash_{\mathsf{MBL}} +B \wedge C$.

A is of the form $\neg B$. If $\sigma^{\mathrm{M}}_{\alpha+1} \models \neg B$, then, by the model-theoretic clauses for negation, $\sigma^{\mathrm{M}}_{\alpha+1} \models\!\!| B$. But then, by the induction hypothesis of the induction on complexity, $\mathsf{TST} \vdash_{\mathsf{MBL}} -B$. Since MBL includes the ($+\neg$I.) rule, it follows that $\mathsf{TST} \vdash_{\mathsf{MBL}} +\neg B$.

A is of the form $\forall x\, F$. If $\sigma^{\mathrm{M}}_{\alpha+1} \models \forall x\, F$, then, by the model-theoretic clauses for the universal quantifier, for all t, $\sigma^{\mathrm{M}} \models F[t/x]$. By the induction hypothesis of the induction on complexity, $\mathsf{TST} \vdash_{\mathsf{MBL}} +F[t/x]$. By the ($+\forall$I.) rule, it follows that $\mathsf{TST} \vdash_{\mathsf{MBL}} +\forall x\, F$.

Limit case. Suppose that $\sigma^{\mathrm{M}}_{\lambda} \models A$ with λ a limit ordinal. Since $\sigma^{\mathrm{M}}_{\lambda} = \bigcup_{\alpha<\lambda} \sigma^{\mathrm{M}}_{\alpha}$, there is some $\beta < \lambda$ such that $\sigma^{\mathrm{M}}_{\lambda} \models A$. By the induction hypothesis of the ordinal induction, $\mathsf{TST} \vdash_{\mathsf{MBL}} +A$. The negative case is analogous. □

Having established that we can axiomatize the fixed point of the Maddy hierarchy, we now provide a general model theory for MBL. An interpretation $\mathcal{I}$ consists of a set of objects $\mathcal{D}$ and an interpretation function $\mathcal{F}$. The function $\mathcal{F}$ assigns elements of $\mathcal{D}$ to closed terms and an EA-pair σ to η where σ^+ and σ^- consist of ordered pairs of elements of $\mathcal{D}$. The satisfaction and anti-satisfaction clauses for identity, the connectives and the universal quantifier are as above. An interpretation $\mathcal{I}$ assigning an EA-pair σ is then said to be η-admissible if, for all A, $\mathcal{I} \models Ft$ just in case $\langle \mathcal{F}(t), \mathcal{F}(\hat{x}F)\rangle \in \sigma^+$, and $\mathcal{I} \models\!\!| Ft$ just in case $\langle \mathcal{F}(t), \mathcal{F}(\hat{x}F)\rangle \in \sigma^-$.

We can now show that MBL is sound and complete with respect to the class of η-admissible models. For soundness, the argument used in Theorem 4.2 above suffices, except that we now use the restriction to η-admissible models to establish the soundness of the η-rules. For completeness, we first need a few additional definitions. We say that a set of formulae Γ is *bilaterally prime* (*b-prime* for short) just in case if $-A \wedge B \in \Gamma$ then either $-A \in \Gamma$ or $-B \in \Gamma$. Moreover, for a given deductive system S, we say that a set Γ is *bilaterally consistent*$_S$ (*b-consistent*$_S$ for short) if for no A is it the case that both $\Gamma \vdash_{\mathsf{S}} +A$ and $\Gamma \vdash_{\mathsf{S}} -A$. We can then prove the following model existence result:

Lemma 4.5. *Let Γ be a b-consistent*$_{\mathsf{MBL}}$ *set of signed sentences and let Γ^* be the extension of its closure under derivability in* MBL *to a b-prime set. Then there is an interpretation $\mathcal{I}$ such that $+A \in \Gamma^*$ iff $\mathcal{I} \models A$, and $-A \in \Gamma^*$ iff $\mathcal{I} \models\!\!| A$.*

Proof. Let $\mathcal{I} = \langle \mathcal{D}, \mathcal{F}\rangle$ be the canonical term model for Γ^* and set $\mathcal{F}(\eta) = \langle \tau^+, \tau^-\rangle$ where τ^+ is $\{\langle \bar{a}, \bar{b}\rangle \mid +a\eta b \in \Gamma^*\}$ and τ^- is $\{\langle \bar{a}, \bar{b}\rangle \mid -a\eta b \in \Gamma^*\}$. Then, for all A, we have that $+A \in \Gamma^*$ iff $\mathcal{I} \models A$, and $-A \in \Gamma^*$ iff $\mathcal{I} \models\!\!| A$. The proof is by induction on the complexity of A.

A is of the form $a\eta b$. $+a\eta b \in \Gamma^*$ iff $\langle \bar{a}, \bar{b}\rangle \in \tau^+$ iff $\mathcal{I} \models a\eta b$. The case of $-a\eta b$ is analogous.

A is of the form $a = b$. $+a = b \in \Gamma^*$ iff $\bar{a} = \bar{b}$ iff $\mathcal{I} \models a = b$. The case of $-a = b$ is analogous.

A is of the form $B \wedge C$. $+B \wedge C \in \Gamma^*$ iff $+B \in \Gamma^*$ and $+C \in \Gamma^*$ (since Γ^* is MBL-closed) iff $\mathcal{I} \models B$ and $\mathcal{I} \models C$ (by the induction hypothesis) iff $\mathcal{I} \models B \wedge C$. Similarly, $-B \wedge C \in \Gamma^*$ iff $-B \in \Gamma^*$ or $-C \in \Gamma^*$ (since Γ^* is b-prime) iff $\mathcal{I} \models\!\!| B$ or $\mathcal{I} \models\!\!| C$ (by the induction hypothesis) iff $\mathcal{I} \models\!\!| B \wedge C$.

A is of the form $\neg B$. $+\neg B \in \Gamma^*$ iff $-B \in \Gamma^*$ (since Γ^* is MBL-closed) iff $\mathcal{I} \models\!\!| B$ (by the induction hypothesis) iff $\mathcal{I} \models \neg B$ (by the model-theoretic clauses for negation). Similarly, $-\neg B \in \Gamma^*$ iff $+B \in \Gamma^*$ iff $\mathcal{I} \models B$ iff $\mathcal{I} \models\!\!| \neg B$.

A is of the form $\forall x\, F$. $+\forall x\, F \in \Gamma^*$ iff for every $t \in T^*$, $F[t/x] \in \Gamma^*$ (since Γ^* is MBL-closed) iff $\mathcal{I} \models Ft$ (by the induction hypothesis) iff $\mathcal{I} \models \forall x\, F$ (by the fact that there is a term for every set and class). Similarly, $-\forall x\, F \in \Gamma$ iff $F[t/x] \in$ MBL for every $t \in T^*$ (since Γ^* is MBL-closed and MBL includes the ($-\forall$E.) rule) iff $\mathcal{I} \models\!\!| Ft$ for every $t \in T^*$ iff $\mathcal{I} \models\!\!| \forall x\, F$. □

We are now in a position to prove completeness. To make the statement of the theorem simpler, it will be useful to extend our model-theoretic notion to cover the signed language. If A is a sentence, we write $\mathcal{I} \models_B +A$ wherever $\mathcal{I} \models A$, and $\mathcal{I} \models_B -A$ wherever $\mathcal{I} \models\!\!| A$. If φ is a signed sentence, we then write $\Gamma \models_B \varphi$ just in case if, for all η-admissible $\mathcal{I}$, if $\mathcal{I} \models_B \psi$ for all $\psi \in \Gamma$, then $\mathcal{I} \models_B \varphi$.

Theorem 4.6. *Let Γ be a set of signed sentences and φ a signed sentence. If $\Gamma \models_B \varphi$, then $\Gamma \vdash_{\mathsf{MBL}} \varphi$.*

Proof. Suppose that $\Gamma \nvdash_{\mathsf{MBL}} \varphi$. It follows that Γ is b-consistent$_{\mathsf{MBL}}$ and so, by Lemma 4.5, there is an interpretation $\mathcal{I}$ and a set of signed sentences Γ^* such that $\mathcal{I} \models A$ iff $+A \in \Gamma^*$, and such that $\mathcal{I} \models\!\!\!| A$ iff $-A \in \Gamma^*$. Since $\Gamma \subseteq \Gamma^*$, clearly, $\mathcal{I} \models_B \psi$ for all $\psi \in \Gamma$. Moreover, $\mathcal{I} \not\models_B \varphi$ since $\varphi \notin \Gamma^*$. So $\Gamma \not\models_B \varphi$. □

In her original paper, Maddy introduces a notation intended to allow us to talk about the situation in which an EA-pair σ does not decide a certain membership fact. Using our notation, we can similarly define:

- $\sigma \perp\!\!\!\perp A$ iff neither $\sigma \models A$ nor $\sigma \models\!\!\!| A$.

With this notation, we can now write that $\sigma^M \perp\!\!\!\perp \hat{x}(x \not\in x)\eta\hat{x}(x \not\in x)$, which expresses model-theoretically the fact that Maddy theory's of sets and classes is agnostic over whether the Russell class belongs to itself. We cannot yet express this fact proof-theoretically. We can do so by moving from a bilateral to a trilateral setting and augmenting the expressive power of MBL with an additional force marker ?, denoting the speech act expressing agnosticism about a certain matter. This speech act is governed by the following coordination principles:

$$\text{(Trilateral Explosion}_1\text{)}\ \frac{+A \qquad ?A}{\varphi} \qquad \text{(Trilateral Explosion}_2\text{)}\ \frac{-A \qquad ?A}{\varphi}$$

$$\text{(Trilateral Excluded Fourth)}\ \frac{\begin{array}{ccc}[+A] & [?A] & [-A] \\ \vdots & \vdots & \vdots \\ \varphi & \varphi & \varphi\end{array}}{\varphi}$$

The Trilateral Explosion$_i$ principles state that it is incoherent to express agnosticism towards a subject matter while at the same time expressing belief or disbelief towards that very same subject matter. Trilateral Excluded Fourth captures the idea that there are three attitudes that can be taken towards a certain content: belief, disbelief or agnosticism. An alternative, and formally equivalent route, would be to take the agnostic attitude to be expressible by weakly asserting (Incurvati and Schlöder, 2019) and weakly rejecting (Incurvati and Schlöder, 2017) the same content (Ferrari and Incurvati, 2022) and lay down corresponding coordination principles.

Now consider the system obtained by extending MBL with the Trilateral Explosion$_i$ principles and Trilateral Excluded Fourth. It is straightforward to prove in this system that $?\hat{x}(x \not\in x)\eta\hat{x}(x \not\in x)$.[9] Indeed, one can show that the system is sound and complete with respect to the general model theory from the previous section augmented with $\perp\!\!\!\perp$ defined as above.

5. Rampant indeterminacy

Maddy's theory of classes satisfies several of the desiderata we discussed above. It validates the unrestricted comprehension rules and hence implies the existence of big classes such as the universal class (i.e. the class containing everything there is) and the Russell class (i.e. the class of all collections that do not belong to themselves). It takes classes to be real, well-defined entities. And it takes classes to be logical collections, since they are characterized in terms of their defining condition, thereby distinguishing them from sets.

[9] In the alternative route using weak assertion and weak rejection, it would be straightforward to prove that $\hat{x}(x \not\in x)\eta\hat{x}(x \not\in x)$ is both weakly asserted and weakly rejected.

However, there are some serious problems with Maddy's theory, which have been stressed in recent work by Linnebo (2024). As Linnebo puts it, Maddy's 'account is very "gappy": there is a *lot* of indeterminacy' (p. 74, emphasis in the original). But in K_3, a universal generalization is indeterminate just in case at least one of its instances is. As a result, it very difficult for a universal generalization to be determinately true in Maddy's theory of classes. This leads to two specific problems, which Linnebo focuses on.

The first concerns Maddy's attempt to develop a theory of Frege–Russell numbers within her theory of classes. Within such a theory, two collections (sets or classes) are equinumerous if and only if there exists a one-to-one correspondence between their members. It follows that two collections are *not* equinumerous if and only every relation fails to be such a correspondence. This is a generalization, which is then subject to the problem of indeterminacy. As a result, it is very hard to prove that two collections are not equinumerous in Maddy's theory. Indeed, as Maddy (2000, 312) herself proves, if the statement that two non-empty collections are equinumerous is not satisfied in her original theory of classes, then it is neither satisfied nor anti-satisfied. It follows that we cannot even prove that $\{\emptyset\}$ is not equinumerous with $\{\emptyset, \{\emptyset\}\}$. Linnebo takes this 'rampant indeterminacy' to be difficult to accept, and Maddy herself takes the result to be troublesome.

The second problem concerns identity in Maddy's theory. Recall that, for any two terms, the model-theoretic clauses for identity declare their referents to be determinately identical just in case the terms themselves are identical, and determinately different just in case the terms themselves are different. This means, in particular, that if F and G are different formulae, then the classes $\hat{x}F$ and $\hat{x}G$ will also be different: Maddy's theory identifies classes in far too fine-grained a manner and hence fails to satisfy our fourth desideratum on a theory of classes.

To repair the situation, one might try to single out a coarser equivalence relation among collections, in terms of which identity among classes could then be defined. Maddy (2000, 305) considers the option of defining identity among collections in terms of the relation $\simeq$, defined as follows.

Definition 5.1. For $t, t' \in T$, $t \simeq t' \equiv_{\text{def}} \forall z(z\eta t \leftrightarrow z\eta t')$

As Maddy observes, this proposal might appear at first sight to give us what we want, since it identifies coextensive sets and classes, such as $\bar{a}$ and $\hat{x}(x\eta\bar{a})$, and certain classes which would have been declared to be different by the original syntactic definition of identity, such as $\hat{x}(x\eta\bar{a})$ and $\hat{x}(x\eta\bar{a} \wedge x\eta\bar{a})$.

However, says Linnebo, the proposal fails because of the problem of rampant indeterminacy to which universal generalizations are subject. As Maddy (2000, 305) points out, σ^{M} can satisfy $u\eta t \leftrightarrow u\eta t'$ only if it is decided about $u\eta t$ or $u\eta t'$. It follows that the definition of identity as $\simeq$ entails that two classes can only be identical if every membership claim concerning them must be either satisfied or anti-satisfied. Indeed, $t \simeq t$ becomes a way of expressing that t is total: $\sigma^M \models t \simeq t$ just in case $\sigma^M \models \forall z(z\eta t \vee z \not\eta t)$. It follows that Maddy's theory is undecided about whether $\hat{x}(x \not\eta x)$ is identical to $\hat{x}(x \not\eta x \wedge x \not\eta x)$. Indeed, the theory is undecided about whether $\hat{x}(x \not\eta x)$ is identical to itself: under the proposed definition, identity is not even reflexive.

Linnebo considers two possible ways of improving on Maddy's account along broadly Maddian lines. The first, which he has developed in joint work with Leon Horsten (Horsten and Linnebo, 2016), consists in building up not an EA-pair for the membership relation, but rather for the identity predicate, so that we have that $\hat{x}F = \hat{x}G$ if and only if $\forall x(Fx \leftrightarrow Gx)$. The

construction proceeds by adding $\langle \hat{x}F, \hat{x}G \rangle$ to the extension of the identity predicate whenever $\forall x(Fx \leftrightarrow Gx)$ is satisfied, and by adding $\langle \hat{x}F, \hat{x}G \rangle$ to the anti-extension of the identity predicate whenever $\forall x(Fx \leftrightarrow Gx)$ is anti-satisfied. However, as Linnebo points out, the construction is severely limited, in that it only works for classes whose defining formula does not contain occurrences of the membership predicate.

The second attempt Linnebo considers is due to Jönne Kriener (2014). Kriener combines ideas from Maddy and from Horsten and Linnebo, in that he builds up EA-pairs for both the membership relation and the identity predicate. The resulting theory, however, has important limitations with regards to the classes whose existence it admits. For instance, the natural way of defining the singleton class of a given class (that is, as $\hat{x}x = \hat{y}F$) fails. Kriener himself takes his results to cast doubt on the viability of a theory of classes based on a hierarchical approach along Kripkean lines.

Linnebo concludes that the prospects for a Maddian account of classes are dim:

> The picture that emerges is thus one of a failed research program. At the outset, Maddy's idea of building on Kripke's highly successful account of truth seemed very promising. And as we have seen, there are natural ways to transpose the account to the case of classes. Unfortunately, the resulting theories are not very attractive; and the situation does not improve materially on any of the more natural ways to modify Maddy's account. (Linnebo, 2024, 75)

In the remainder of this paper, I will argue that, contrary to appearances, the prospects for a hierarchical approach to classes along broadly Kripkean-Maddian lines are in fact bright. There is a very natural way of modifying Maddy's account so as to solve the problem highlighted by Linnebo.

6. Supervaluations

The rampant indeterminacy observed by Linnebo is indeed a problematic feature of Maddy's theory. The problem is a familiar one: it already affects Kripke's theory of truth and is due to the fact that the underlying logic of Maddy's theory, just like Kripke's, is K3.

A natural strategy to deal with the problem is to adopt a different scheme for handling membership gaps. In his original paper, Kripke (1975, 711–712) already suggested the possibility of handling truth-value gaps using a scheme based on Bas van Fraassen's (1966) notion of supervaluation. van Fraassen originally introduced the idea of supervaluation to handle truth-value gaps generated by the use of empty singular terms. In the subsequent literature, much attention has been devoted to the application of the supervaluational strategy to the case of vagueness (Fine, 1975; Incurvati and Schlöder, 2022; Keefe, 2000).

In the case of truth, the idea is to handle truth-value gaps by assigning a sentence A to the extension (anti-extension) of the truth predicate at the next stage of the construction just in case all admissible expansions of the EA-pair of the truth predicate at the current stage classically satisfy (anti-satisfy) A. Transposing this to the class-theoretic setting, the idea is to handle membership gaps by assigning an ordered pair $\langle t, \hat{x}F \rangle$ consisting of an object and a class to the extension (anti-extension) of the membership predicate at the next stage of the construction just in case all admissible expansions the EA-pair of the membership predicate at the current stage classically satisfy (anti-satisfy) Ft. This gives rise to the following template

for class supervaluational schemes, where the satisfaction relation $\models$ and the antisatisfaction relation $=\!|$ are recursively defined as before, except that we now require that, for every EA-pair σ and for every sentence A, either $\sigma \models A$ or $\sigma =\!| A$:

Definition 6.1 (Supervaluational-jump template)**.** Given an EA-pair $\sigma = \langle \sigma^+, \sigma^- \rangle$, the *supervaluational jump* J_{s} is the operation such that, for all $t \in T^*$,

$$J_{\mathrm{s}}(\sigma^+) = \{\langle t, \hat{x}F \rangle \mid \text{ for all } \tau \text{ that are s-admissible expansions of } \sigma, \tau \models Ft\}, \text{ and}$$
$$J_{\mathrm{s}}(\sigma^-) = \{\langle t, \hat{x}F \rangle \mid \text{ for all } \tau \text{ that are s-admissible expansions of } \sigma, \tau =\!| Ft\}.$$

Different supervaluational schemes are then obtained by specifying when an EA-pair is an admissible expansion of another EA-pair. The first supervaluational scheme considered by Kripke is a straightforward adaptation of van Fraassen's original notion of supervaluation to the case of truth. In particular, the idea is to take an admissible expansion of an EA-pair σ to be one that does not satisfy (anti-satisfy) any sentence that is anti-satisfied (satisfied) by σ. We can proceed in an analogous manner in the case of classes: an admissible expansion is one that respects the already established membership facts.

Definition 6.2 (MvF-admissible expansion)**.** An EA-pair $\tau = \langle \tau^+, \tau^- \rangle$ is a *MvF-admissible expansion* of an EA-pair $\sigma = \langle \sigma^+, \sigma^- \rangle$ iff (i) $\tau \sqsupseteq \sigma$, (ii) $\tau^+ \cap \sigma^- = \emptyset$, and (iii) $\tau^- \cap \sigma^+ = \emptyset$.

By replacing the Maddy jump J_M with the Maddy–van Fraassen jump J_{MvF} in the definition of the Maddy hierarchy, we obtain the *basic supervaluational class hierarchy*. It is easy to check that (still taking $\bar{V}$ to denote V_κ where k is the first inaccessible) the construction of the basic supervaluational class hierarchy reaches a fixed point σ^{MvF}. We can therefore consider σ^{MvF} as giving the extension of η according to the Maddy–van Fraassen theory of classes—the theory of classes obtained by adopting the van Fraassen supervaluational scheme within the context of a theory of classes along broadly Maddian lines.

I now want to present a natural deduction system for the Maddy–van Fraassen theory of classes. It is well known that the addition of the following rule of Bilateral Excluded Middle to the asserted and rejected rules for negation and conjunction of MBL and Bilateral Explosion delivers a bilateral version of classical propositional logic:

$$\text{(Bilateral Excluded Third)}\ \frac{\begin{array}{cc}[+A] & [-A] \\ \vdots & \vdots \\ \varphi & \varphi\end{array}}{\varphi}$$

For the principles of Bilateral Explosion and Bilateral Excluded Third are interderivable with the following coordination principles (see, e.g., del Valle-Inclan, 2023, 382), which are sufficient to provide a sound and complete axiomatization of the classical propositional calculus in the presence of the rules for negation and conjunction of MBL (Incurvati and Schlöder, 2023; Rumfitt, 2000; Smiley, 1996), where '(SR$_i$)' abbreviates 'Smileian reductio$_i$'.

$$\text{(Rejection)}\ \frac{+A \qquad -A}{\bot} \qquad \text{(SR}_1\text{)}\ \frac{\begin{array}{c}[+A] \\ \vdots \\ \bot\end{array}}{-A} \qquad \text{(SR}_2\text{)}\ \frac{\begin{array}{c}[-A] \\ \vdots \\ \bot\end{array}}{+A}$$

Now, a central supervaluational intuition is that classical reasoning is in order when no rules are used that can gender indeterminacy. And in our current context, this indeterminacy can only arise because of the application of the η-rules. This suggests restricting the application of these rules within the coordination principle of Bilateral Excluded Third, in order to obtain a supervaluationistically acceptable version thereof. We thereby obtain the following:

$$\text{(BET*)}\ \frac{\begin{array}{cc}[+A] & [-A] \\ \vdots & \vdots \\ \varphi & \varphi\end{array}}{\varphi}\ \text{if the subderivations of } \varphi \text{ use no } \eta\text{-rules}$$

Let MvFBL be the system obtained by adding (BET*) to MBL and restricting the subderivations in the ($-\wedge$E.) and ($-\forall$E.) rules in the same manner (that is, to subderivations that make no use η-rules). I am now going to prove that analogous results hold between MvFBL and the Maddy–van Fraassen theory of classes as those that obtained between MBL and Maddy's theory of classes.

I begin by proving that, over TST, MvFBL axiomatizes the fixed point of the basic supervaluational hierarchy. I first prove that MvFBL is sound with respect to the consequence relation induced by σ^{MvF}.

Theorem 6.3. *Let Γ be a set of signed sentences and φ a signed sentence. Suppose that for all ψ in Γ, $\sigma^{\text{MvF}} \models A$ if $\psi = +A$, and $\sigma^{\text{MvF}} \nvDash A$ if $\psi = -A$. Then, if $\Gamma \vdash_{\mathsf{MvFBL}} \varphi$, $\sigma^{\text{MvF}} \models B$ if $\varphi = +B$ and $\sigma^{\text{MvF}} \nvDash B$ if $\varphi = -B$.*

Proof. The arguments in the proof of Theorem 4.2 carry over to the present case except for the ($-\wedge$E.) and ($-\forall$E.) rules, the (BET*) principle and the η-rules. For the ($-\wedge$E.) and ($-\forall$E.) rules and the (BET*) principle, it suffices to note that all of their instances are instances of classically valid arguments (see Incurvati and Schlöder, 2017). It remains to check the η-rules. We only cover the positive cases because the negative ones proceed in a completely analogous fashion by replacing the satisfaction relation with the anti-satisfaction one.

For the ($+\eta$I.) rule, suppose that $\sigma^{\text{MvF}} \models Ft$. By monotonicity, this means that for every $\tau \sqsupseteq \sigma^{\text{MvF}}$, $\tau \models Ft$. So, in particular, for every τ which is an MvF-admissible expansion of σ^{MvF}, $\tau \models Ft$. By the definition of the Maddy–van Fraassen supervaluational jump, it follows that $\langle t, \hat{x}F\rangle \in J_{\text{MvF}}(\sigma^{\text{MvF}})$. Since σ^{MvF} is a fixed point, it follows that $\sigma^{\text{MvF}} \models t\eta\hat{x}F$.

For the ($+\eta$E.) rule, suppose that $\sigma^{\text{MvF}} \models t\eta\hat{x}F$. By the definition of J_{MvF}, this means that for every τ which is an MvF-admissible expansion of σ^{MvF}, $\tau \models Ft$. Since σ^{MvF} is an admissible expansion of itself, we have, in particular, that $\sigma^{\text{MvF}} \models Ft$. □

As in the case of Maddy's theory of classes, by construction we have that $\sigma^{\text{MvF}} \models A$ for all $+A \in \mathsf{TST}$ and $\sigma^{\text{MvF}} \nvDash A$ for all $-A \in \mathsf{TST}$. Hence, Theorem 6.3 immediately yields that, over TST, MvFBL is sound with respect to σ^{MvF}.

Theorem 6.4. *For every sentence A, if $\mathsf{TST} \vdash_{\mathsf{MvFBL}} +A$, then $\sigma^{\text{MvF}} \models A$, and if $\mathsf{TST} \vdash_{\mathsf{MvFBL}} -A$, then $\sigma^{\text{MvF}} \nvDash A$.*

I now turn to the proof that, over TST, MvFBL is complete with respect to σ^{MvF}. For a given deductive system S, we say that Γ is maximally b-consistent$_{\mathsf{S}}$ if it is b-consistent$_{\mathsf{S}}$ and either

$+A \in \Gamma$ or $-A \in \Gamma$. Moreover, for every system S, we let S^* be S without the η-rules. Using Zorn's Lemma, we can prove the following:[10]

Lemma 6.5. *Every b-consistent$_{\mathsf{MvFBL}^*}$ set of signed sentences can be extended to a maximally b-consistent$_{\mathsf{MvFBL}^*}$ set.*

We can now prove that every maximally b-consistent$_{\mathsf{MvFBL}^*}$ extension of TST closed under derivability in MvFBL has a model:

Theorem 6.6. *Suppose that $\Gamma \supseteq \mathsf{TST}$ is a maximally b-consistent$_{\mathsf{MvFBL}^*}$ set of signed sentences closed under derivability in MvFBL, and let τ be the EA-pair $\langle \tau^+, \tau^- \rangle$, where τ^+ is $\{\langle \bar{a}, \bar{b} \rangle \mid +a\eta b \in \Gamma\}$ and τ^- is $\{\langle \bar{a}, \bar{b} \rangle \mid -a\eta b \in \Gamma\}$. Then, for all sentences A, we have that $+A \in \Gamma$ iff $\tau \models A$, and $-A \in \Gamma$ iff $\tau \rightleftharpoons A$.*

Proof. By induction on the complexity of A.

A is of the form $t\eta t'$. By definition of τ.

A is of the form $t = t'$. Suppose $+t = t' \in \Gamma$ but $\tau \not\models t = t'$. Then t is not the same term as t' and hence, since Γ is closed under MvFBL-derivability and MvFBL includes the ($-=$I.) rule, $-t \neq t' \in \Gamma$. But this contradicts the assumption that Γ is b-consistent$_{\mathsf{MvFBL}^*}$. For the other direction, if $\tau \models t = t'$, then, by the model-theoretic clauses for identity, t is the same term as t'. But then, $+t = t' \in \Gamma$, since MvFBL includes the ($+=$I.) rule. The negative case is analogous.

A is of the form $B \wedge C$. Suppose $+B \wedge C \in \Gamma$. Then $+B \in \Gamma$ and $+C \in \Gamma$, since Γ is closed under MvFBL-derivability and MvFBL includes the ($+\wedge$I.) rule. By the induction hypothesis, $\tau \models B$ and $\tau \models C$. By the model-theoretic clauses for conjunction, $\tau \models B \wedge C$. The reverse direction is analogous. For the negative case, suppose $-B \wedge C \in \Gamma$. Then $-B \in \Gamma$ or $-C \in \Gamma$, since Γ is maximally b-consistent$_{\mathsf{MvFBL}^*}$. By the induction hypothesis, $\tau \rightleftharpoons B$ or $\tau \rightleftharpoons C$. By the model-theoretic clauses for conjunction, $\tau \rightleftharpoons B \wedge C$. For the other direction, suppose $\tau \rightleftharpoons B \wedge C$. Then $\tau \rightleftharpoons B$ or $\tau \rightleftharpoons C$. By the induction hypothesis, $-B \in \Gamma$ or $-C \in \Gamma$. Since Γ is closed under MvFBL-derivability and MvFBL includes the ($-\wedge$I.) rule, it follows that $-B \wedge C \in \Gamma$.

A is of the form $\neg B$. If $+\neg B \in \Gamma$, then $-B \in \Gamma$, since MvFBL includes the ($+\neg$I.) rule. By the induction hypothesis, it follows that $\tau \rightleftharpoons B$ and hence that $\tau \models \neg B$. The reverse direction is analogous, and so are the negative cases.

A is of the form $\forall x\, F$. If $+\forall x\, F \in \Gamma$, then, for all $t \in T^*$, $+F[t/x] \in \Gamma$, since MvFBL includes the ($+\forall$I.) rule. By the induction hypothesis, it follows that, for all $t \in T^*$, $\tau \models F[t/x]$, and, by the model-theoretic clauses for the universal quantifier, that $\tau \models \forall x\, F$. The reverse direction is analogous, and so are the negative cases. □

We are now in a position to prove completeness.

Theorem 6.7. *For every sentence A, if $\sigma^{\mathrm{MvF}} \models A$, then $\mathsf{TST} \vdash_{\mathsf{MvFBL}} +A$, and if $\sigma^{\mathrm{MvF}} \rightleftharpoons A$, then $\mathsf{TST} \vdash_{\mathsf{MvFBL}} -A$.*

Proof. I prove that if $\sigma_\alpha^{\mathrm{MvF}} \models A$, then $\mathsf{TST} \vdash_{\mathsf{MvFBL}} +A$, and if $\sigma_\alpha^{\mathrm{MvF}} \rightleftharpoons A$, then $\mathsf{TST} \vdash_{\mathsf{MvFBL}} -A$. The proof proceeds exactly like the proof of Theorem 4.4, except for the case in which A is of the form $t\eta t'$ in the induction on the complexity of A in the inductive step of the ordinal induction.

[10] In the context of a countable language, it would be possible to construct the maximally b-consistent set using a step-by-step procedure. See Incurvati and Schlöder, 2023, 568–569 for an application of the method in the theory of truth.

So suppose that $\sigma^{\mathrm{MvF}}_{\alpha+1} \models t\eta t'$. If t' is a set term or $\bar{V}$, we have that $\mathsf{TST} \vdash_{\mathsf{MvFBL}} +t\eta t'$ by the fact that $\sigma^{\mathrm{MvF}}_{\alpha+1} \sqsupseteq \sigma^{\mathrm{MvF}}_{0}$. So let t' be of the form $\hat{x}F$ for some F. Now suppose that $\mathsf{TST} \nvdash_{\mathsf{MvFBL}} +t\eta\hat{x}F$. This means that the deductive closure of $\mathsf{TST} \cup -t\eta\hat{x}F$ under MvFBL-derivability is b-consistent$_{\mathsf{MvFBL}^*}$ and so it has a maximally b-consistent$_{\mathsf{MvFBL}^*}$ extension Γ. Let $\tau = \langle \tau^+, \tau^- \rangle$, where τ^+ is $\{\langle \bar{a}, \bar{b} \rangle \mid +a\eta b \in \Gamma\}$ and τ^- is $\{\langle \bar{a}, \bar{b} \rangle \mid -a\eta b \in \Gamma\}$. I now show that τ is an MvF-admissible expansion of $\sigma^{\mathrm{MvF}}_{\alpha}$.

First, suppose that there was a formula $t\eta\hat{x}G$ such that $\sigma^{\mathrm{MvF}}_{\alpha} \models t\eta\hat{x}G$ but $\tau \not\models t\eta\hat{x}G$. By the induction hypothesis, it follows that $\mathsf{TST} \vdash_{\mathsf{MvFBL}} +t\eta\hat{x}G$. But then, by construction of τ, $\tau \models t\eta\hat{x}G$, which contradicts the original supposition. Hence, (i) $\tau \sqsupseteq \sigma^{\mathrm{MvF}}_{\alpha}$. Next, suppose that there is a $\langle t, \hat{x}G \rangle$ such that $\langle t, \hat{x}G \rangle \in \tau^+$ and $\langle t, \hat{x}G \rangle \in \sigma^{\mathrm{MvF},-}_{\alpha}$. By definition of τ, it follows that $+t\eta\hat{x}G \in \Gamma$. And by the induction hypothesis, it follows that $\mathsf{TST} \vdash_{\mathsf{MvFBL}} -t\eta\hat{x}G$ and so that $-t\eta\hat{x}G \in \Gamma$. But since Γ is b-consistent$_{\mathsf{MvFBL}^*}$, we have that $+t\eta\hat{x}G \notin \Gamma$, which by construction of τ means that $\langle t, \hat{x}G \rangle \notin \tau^+$. This contradicts our original supposition. Hence, (ii) $\tau^+ \cap \sigma^{\mathrm{MvF},-}_{\alpha} = \emptyset$. A similar reasoning shows that (iii) $\tau^- \cap \sigma^{\mathrm{MvF},+}_{\alpha} = \emptyset$.

By supposition, we have that $-t\eta\hat{x}F \in \Gamma$. Since Γ is closed under MvFBL-derivability, this means that $-Ft \in \Gamma$ too. By Theorem 6.6 it then follows that $\tau \nRightarrow Ft$. So there is an MvF-admissible expansion of $\sigma^{\mathrm{MvF}}_{\alpha}$ which does not satisfy Ft. This contradicts the supposition that $\sigma^{\mathrm{MvF}}_{\alpha+1} \models t\eta\hat{x}F$. So if $\sigma^{\mathrm{MvF}}_{\alpha+1} \models t\eta\hat{x}F$, then $\mathsf{TST} \vdash_{\mathsf{MvFBL}} +t\eta\hat{x}F$. □

I now provide a general model theory for MvFBL. An interpretation $\mathcal{I}$ consists of a set of objects $\mathcal{D}$ and an interpretation function $\mathcal{F}$. The function $\mathcal{F}$ assigns elements of $\mathcal{D}$ to closed terms and an EA-pair σ to η where σ^+ and σ^- consists of ordered pairs of elements of $\mathcal{D}$. The satisfaction and anti-satisfaction clauses for identity, the connectives and the universal quantifier are as above. We then need to extend the definition of model-theoretic consequence to cover the signed language. If A is a sentence, we write $\mathcal{I}_\sigma \models_S +A$ wherever, for all τ that are MvF-admissible expansions of σ, $\mathcal{I}_\tau \models A$, and $\mathcal{I}_\sigma \models_S -A$ wherever, for all for all τ that are MvF-admissible expansions of σ, $\mathcal{I}_\tau \nRightarrow A$. If Γ is a set of signed sentences and φ a signed sentence, we then write $\Gamma \models_S \varphi$ just in case if, for all $\mathcal{I}$, if $\mathcal{I} \models_S \psi$ for all $\psi \in \Gamma$, then $\mathcal{I} \models_S \varphi$.

We can now show that MvFBL is sound and complete with respect to the class of η-admissible models. For soundness, the argument used in Theorem 6.3 above suffices, except that we now use the restriction to η-admissible models to establish the soundness of the η-rules. For completeness, we prove the following model existence result:

Lemma 6.8. *Let Γ be a b-consistent$_{\mathsf{MvFBL}^*}$ set of signed sentences and let Γ^* be a maximally b-consistent$_{\mathsf{MvFBL}^*}$ extension of its closure. Then there is an interpretation $\mathcal{I}$ such that $+A \in \Gamma^*$ iff $\mathcal{I} \models A$, and $-A \in \Gamma^*$ iff $\mathcal{I} \nRightarrow A$.*

Proof. Let $\mathcal{I} = \langle \mathcal{D}, \mathcal{F} \rangle$ be the canonical term model for Γ^* and set $\mathcal{F}(\eta) = \langle \tau^+, \tau^- \rangle$ where τ^+ is $\{\langle \bar{a}, \bar{b} \rangle \mid +a\eta b \in \Gamma^*\}$ and τ^- is $\{\langle \bar{a}, \bar{b} \rangle \mid -a\eta b \in \Gamma^*\}$. Then, for all A, we have that $+A \in \Gamma^*$ iff $\mathcal{I} \models A$, and $-A \in \Gamma^*$ iff $\mathcal{I} \nRightarrow A$. The proof is by induction on the complexity of A.

A is of the form $a\eta b$. $+a\eta b \in \Gamma^*$ iff $\langle \bar{a}, \bar{b} \rangle \in \tau^+$ iff $\mathcal{I} \models a\eta b$. The case of $-a\eta b$ is analogous.

A is of the form $a = b$. $+a = b \in \Gamma^*$ iff $\bar{a} = \bar{b}$ iff $\mathcal{I} \models a = b$. The case of $-a = b$ is analogous.

A is of the form $B \wedge C$. $+B \wedge C \in \Gamma^*$ iff $+B \in \Gamma^*$ and $+C \in \Gamma^*$ (since Γ^* is MBL-closed) iff $\mathcal{I} \models B$ and $\mathcal{I} \models C$ (by the induction hypothesis) iff $\mathcal{I} \models B \wedge C$. Similarly, $-B \wedge C \in \Gamma^*$ iff $-B \in \Gamma^*$ or $-C \in \Gamma^*$ (since Γ^* is maximally b-consistent$_{\mathsf{MvFBL}^*}$) iff $\mathcal{I} \nRightarrow B$ or $\mathcal{I} \nRightarrow C$ (by the induction hypothesis) iff $\mathcal{I} \nRightarrow B \wedge C$.

A is of the form $\neg B$. $+\neg B \in \Gamma^*$ iff $-B \in \Gamma^*$ (since Γ^* is MvFBL-closed) iff $\mathcal{I} \neq\!|\, B$ (by the induction hypothesis) iff $\mathcal{I} \models \neg B$ (by the model-theoretic clauses for negation). Similarly, $-\neg B \in \Gamma^*$ iff $+B \in \Gamma^*$ iff $\mathcal{I} \models B$ iff $\mathcal{I} \neq\!|\, \neg B$.

A is of the form $\forall x\, F$. $+\forall x\, F \in \Gamma^*$ iff for every $t \in T^*$, $F[t/x] \in \Gamma^*$ (since Γ^* is MvFBL-closed) iff $\mathcal{I} \models Ft$ (by the induction hypothesis) iff $\mathcal{I} \models \forall x\, F$ (by the fact that there is a term for every set and class). Similarly, $-\forall x\, F \in \Gamma$ iff $F[t/x] \in$ MvFBL for every $t \in T^*$ (since Γ^* is MvFBL-closed and MvFBL includes the ($-\forall$E.) rule) iff $\mathcal{I} \neq\!|\, Ft$ for every $t \in T^*$ iff $\mathcal{I} \neq\!|\, \forall x\, F$. □

We can now establish completeness.

Theorem 6.9. *Let* Γ *be a set of signed sentences and* φ *a signed sentence. If* $\Gamma \models_S \varphi$, *then* $\Gamma \vdash_{\mathsf{MvFBL}} \varphi$.

Proof. Suppose that $\Gamma \nvdash_{\mathsf{MvFBL}} \varphi$. It follows that Γ is b-consistent$_{\mathsf{MvFBL}^*}$ and so, by Lemma 6.8, there is an interpretation $\mathcal{I}_\sigma$ and a set of signed sentences Γ^* such that $\mathcal{I}_\sigma \models A$ iff $+A \in \Gamma^*$, and such that $\mathcal{I}_\sigma \neq\!|\, A$ iff $-A \in \Gamma^*$. Let $\mathcal{I}'_\tau$ be the interpretation obtained by taking the set of all these $\mathcal{I}_\sigma$ and defining τ as in Theorem (just above). It is easy to check that τ is maximal superset of the set of all admissible expansions of σ. Hence, using Lemma 6.8, $\mathcal{I}_\tau \models_S \psi$ for all $\psi \in \Gamma^*$ and hence for all $\psi \in \Gamma$ (since $\Gamma \subseteq \Gamma'$). Moreover, $\mathcal{I}_\tau \not\models_S \varphi$ since $\varphi \notin \Gamma^*$. So $\Gamma \vdash_{\mathsf{MvFBL}} \varphi$. □

7. Identity

The Maddy–van Fraassen approach goes a long way towards addressing the problem of rampant indeterminacy. For the supervaluational character of the theory means that more generalizations will be declared true than in Maddy's original theory, due to the fact that instances of those generalizations are satisfied or antisatisfied that were not in MBL.

For example, it is indeterminate whether the Russell class belongs to itself: MvFBL proves neither $+\hat{x}(x \not\eta x)\eta\hat{x}(x \not\eta x)$ nor $+\hat{x}(x \not\eta x)\ \not\eta\hat{x}(x \not\eta x)$. Indeed, when augmented with a force marker for the speech act expressing agnosticism and corresponding coordination principles (as we did in the case of MBL), MvFBL proves $?\hat{x}(x \not\eta x)\eta\hat{x}(x \not\eta x)$. Nonetheless, MvFBL proves that every thing either belongs or does not belong to the Russell class, that is $+\forall y(y\eta\hat{x}(x \not\eta x) \vee y\ \not\eta\hat{x}(x \not\eta x))$.

The supervaluational character of the Maddy–van Fraassen approach also opens up the way to a Frege–Russell treatment of cardinality. As we saw above, Maddy had proved that if the statement that two non-empty collections are equinumerous is not satisfied in her original theory of classes, then it is neither satisfied nor anti-satisfied. To understand the proof, it will be helpful to state in full the Frege–Russell definition of equinumerosity adopted by Maddy.

Definition 7.1. For $t, t' \in T$, $t \approx t'$ abbreviates

$$\exists z(\forall u \forall v \forall w((\langle u, v\rangle \eta z \wedge \langle u, w\rangle \eta z \supset v = w) \wedge (\langle u, v\rangle \eta z \wedge \langle w, v\rangle \eta z \supset u = w)) \wedge$$
$$\forall u((u \eta t \supset \exists v(v \eta t' \wedge \langle u, v\rangle \eta z)) \wedge (u \eta t' \supset \exists v(v \eta t \wedge \langle v, u\rangle \eta z))))$$

Maddy's proof then goes as follows. To establish that $t \not\approx t'$, we need to show that every member of T^* falsifies one of the conjuncts of the definition of equinumerosity. Maddy then shows that if t and t' are non-empty, this cannot be. A crucial step in the proof is that the two conjuncts of

the first clause, namely

$$((\langle u, v\rangle \eta z \wedge \langle u, w\rangle \eta z \supset v = w) \wedge (\langle u, v\rangle \eta z \wedge \langle w, v\rangle \eta z \supset u = w))$$

can never be falsified because there are classes such that, for any collection, the theory does not decide whether the collection belongs to them. (For instance, for every t, $\sigma^M \perp\!\!\!\perp t\eta\hat{x}(x \not\eta y)$.) However, this step fails in the Maddy–van Fraassen theory of classes, since a material conditional $A \to B$ can be false even though both A and B are indeterminate. Similar considerations apply to the reasoning Maddy employs to establish that the two conjuncts of the second clause can never be falsified.

The Maddy–van Fraassen approach does not go all the way towards addressing the problem of rampant indeterminacy, however. Recall that Linnebo's second problem concerned identity and in particular the fact that identity among classes is too fine-grained on Maddy's theory, so that whenever F and G are different formulae, then $\hat{x}F$ and $\hat{x}G$ are different classes. As we saw, a natural option to deal with the problem, one already considered by Maddy, is to define identity in terms of the equivalence relation $\simeq$, where $t \simeq t' \equiv_{\text{def}} \forall z(z\eta t \leftrightarrow z\eta t')$. While promising, this definition failed to even be reflexive because of the problem of rampant indeterminacy. In the Maddy–van Fraassen theory, by contrast, the relation behaves structurally as we want. In particular, $\sigma^{\text{MvF}} \models \forall z(z\eta t \leftrightarrow z\eta t)$ and so $\sigma^{\text{MvF}} \models t \simeq t$. Indeed, it is easy to check that $\simeq$ is an equivalence relation in the Maddy–van Fraassen theory of classes, as desired.

Nonetheless, the problem of identity is far from being solved. For we still have that $\sigma^{\text{MvF}} \perp\!\!\!\perp \forall z(z\eta\hat{x}(x \not\eta x) \leftrightarrow z\eta\hat{x}(x \not\eta x \wedge x \not\eta x))$. It follows that, like Maddy's theory, the Maddy–van Fraassen theory is undecided about whether $\hat{x}(x \not\eta x)$ is identical to $\hat{x}(x \not\eta x \wedge x \not\eta x)$. The reason, essentially, is that, in the definition of the Maddy–van Fraassen jump, we are supervaluating over all expansions consistent with what has already been established about the extension and anti-extension of the membership relation. These expansions include ones in which $\hat{x}(x \not\eta x)$ and $\hat{x}(x \not\eta x \wedge x \not\eta x)$ have different membership constituencies. Should the identity problem persist, Linnebo's negative assessment of the prospects for a hierarchical approach to classes based on a Kripke-style construction would after all be justified.

From the way I have introduced it, one might form the impression that the identity problem only besets Maddy's original theory and its supervaluational cousin. This impression would be mistaken. For the problem is really a problem about how to treat identity within the context of a naïve theory of classes—that is, a theory of classes that validates the unrestricted η-rules.

The issue is brought into sharp relief by Restall (2010), who presents a result that spells trouble for Hartry Field's (2008) theory of properties and indeed any naïve theory of properties or classes. Restall's result, which generalizes a theorem of Roland Hinnion (Hinnion and Libert, 2003), shows that there are limits to how finely classes can be identified. In particular, suppose that we take identity among classes to be governed by the following introduction and elimination rules:

$$(+{=}\,\text{I.})\ \frac{\begin{array}{cc}[+t\eta\hat{x}F] & [+t\eta\hat{x}G] \\ \vdots & \vdots \\ +t\eta\hat{x}G & +t\eta\hat{x}F\end{array}}{+\hat{x}F = \hat{x}G}\ \begin{array}{l}\text{if the subderivations of } +t\eta\hat{x}G \text{ and } +t\eta\hat{x}F \\ \text{use no side premisses}\end{array}$$

$$(+{=}\,\text{E.})\ \frac{+t\eta\hat{x}F \qquad +\hat{x}F = \hat{x}G}{+t\eta\hat{x}G}$$

Then, we can derive a contradiction using only the asserted η-rules and assuming the existence of a sentence from whose assertion anything follows (which we do have in MBL and hence MvFBL, since both system includes the principle of Bilateral Explosion).[11] Essentially, what is happening is that the resources of class theory allow us to derive a Curry-style result without using a conditional, whose logical complexity is simulated by using the class-term operator $\hat{}$.

The difficulties surrounding extensionality in the context of a naïve theory of classes have long been known (Gilmore, 1974; Grišin, 1982; White, 1979). One might however have been tempted to set those difficulties aside on the grounds that full-blown extensionality for classes, understood as logical collections, might not be desirable anyway: while the set of chordates and the set of renates, to use Quine's (1951) famous example, are the same, the corresponding classes or properties are different. Restall's result makes it clear that this reaction would be hasty. For the identity rules used in his derivation do not require classes to be extensional, since the subderivation in the introduction rule is restricted so as to rule out side premisses. More than mere extensional equivalence is needed for two classes to be declared identical. Indeed, Restall takes ($+=$ I.) to encode the idea that (rephrasing things in our framework) if asserting $t\eta\hat{x}F$ commits one to asserting $t\eta\hat{x}G$ on purely logical grounds, and if asserting $t\eta\hat{x}G$ commits one to asserting $t\eta\hat{x}F$ on purely logical grounds, then one is already committed to asserting that $\hat{x}F$ and $\hat{x}G$ are identical. Restall goes on to argue that this intensional criterion of identity of classes is hard to reject:

> To reject [($+=$ I.)] is to reject a coarse account of properties. It is to not only accept a finer individuation of properties where logically equivalent statements pick out distinct properties, but to hold that any attempt at defining coarse properties in terms of fine ones must fail. But at what point does that construction break down? It seems like a straightforward construction of equivalence classes or their representatives. To reject [($+=$ I.)] is also to leave open the vexed question of what the identity conditions for properties can be, and should be. Just what is a property that requires that they be more finely individuated than logic requires? Field's model construction doles out properties in exactly the measure of our language. But it would be bizarre to think that this is an adequate picture of properties. How convenient it would be if properties fit our language like hand to glove? But which language? My language now? Or yours? How are we to get to an adequate understanding of the properties picked out by this naïve theory of properties? (Restall, 2010, pp. 442–443)

Restall is certainly correct that an identity criterion for classes which closely tracks their syntax is unsatisfactory. This holds for the criterion of identity which can be extracted from the model-theoretic construction of Field (2008), which is the focus of Restall's paper, as much as it does for Maddy's original definition of identity.

It is less clear, however, that '[t]o reject [($+=$ I.)] is to reject a coarse account of properties' (or classes). In his response (Field, 2010), Field interprets Restall's ($+=$ I.) as proof-theoretically encoding the model-theoretic principle that if $t\eta\hat{x}F \models_w t\eta\hat{x}G$ and $t\eta\hat{x}G \models_w t\eta\hat{x}F$, then $\models_w \hat{x}F = \hat{x}G$ (where $A \models_w B$ is *weak entailment*: in every model in which A is satisfied, so is B). But, Field

[11] Restall presents his derivation in a sequent calculus setting, which makes explicit the structural assumptions. In that context, the derivation uses the rules of Reflexivity and Cut.

says, this is implausible. If we let L be a Liar sentence, and F be '$L \wedge x$is a kangaroo' and G be '$L \wedge x$is a cockroach', then $t\eta\hat{x}F \models_w t\eta\hat{x}G$ and $t\eta\hat{x}G \models_w t\eta\hat{x}F$ in Field's theory of truth, but we would not want to conclude that $\hat{x}F = \hat{x}G$.

In more recent work, Restall has returned to the issue in collaboration with Shawn Standefer and Rohan French (Standefer et al., 2020). They point out that Restall's result is, first and foremost, a proof-theoretic result, based on the proof-theoretic principle that if I can derive $t\eta\hat{x}G$ from $t\eta\hat{x}F$ (appealing to other assumptions) and *vice versa*, then I can derive $\hat{x}F = \hat{x}G$. Field assumes that if I can derive B from A appealing to no other assumptions, all I learn is that A weakly entails B. But perhaps I learn more. Perhaps I learn that A strongly entails B, where A strongly entails B just in case $\models_w A \to B$. Given the failure of the Deduction Theorem for the conditional Field uses and as suggested by the terminology, weak entailment does not imply strong entailment.

Now, as Standefer et al. (2020) note, it is hard to tell what is needed proof-theoretically in order to establish a strong entailment, since Field has not provided a proof theory for his theory of truth, but only a model theory. We can do so, however, for the Maddy–van Fraassen theory of classes. To this end, it will be helpful to return to the definition of identity as $\simeq$. It is easy to see that, in the context of the supervaluationist setting of the Maddy–van Fraassen theory, this definition sanctions the following rules:

$$(+\simeq \text{I.})\ \frac{\begin{array}{cc}[+t\eta\hat{x}F] & [+t\eta\hat{x}G] \\ \vdots & \vdots \\ +t\eta\hat{x}G & +t\eta\hat{x}F\end{array}}{+\hat{x}F \simeq \hat{x}G}\ \text{if the subderivations of } +t\eta\hat{x}G \text{ and } +t\eta\hat{x}F \text{ use no } \eta\text{-rules}$$

$$(+\simeq \text{E.})\ \frac{+t\eta\hat{x}F \qquad +\hat{x}F \simeq \hat{x}G}{+t\eta\hat{x}G}$$

Thus, there is at least a sense, brought out by the proof theory, in which defining identity as $\simeq$ in the context of the Maddy–van Fraassen theory is both more permissive and more restrictive than the definition of identity for classes/properties proposed by Restall. It is more permissive in that it does allow side premisses in the subderivations of the introduction rule for asserted identity. It is more restrictive in that it does not allow the use of the η-rules in the subderivations. In general, to derive $+B$ from $+A$ without using any other assumptions does not suffice in MvFBL to establish that A strongly entails B where strong entailment is defined using the material conditional (that is, $\sigma^{\text{MvF}} \models A \to B$). What is rather needed is that the derivation from $+A$ to $+B$ does not use the η-rules.[12]

Using our proof-theoretic resources, we can then clearly explain what more is needed in our setting to establish a strong entailment. However, we still have not solved the problem that the proposed criterion of identity among classes appears far too restrictive. For it is still the case that we cannot conclude in MvFBL that $\hat{x}(x \not\eta x)$ and $\hat{x}(x \not\eta x \wedge x \not\eta x)$ are identical. So one might suspect that Restall was after all right that to reject $(+=\text{I.})$ is to reject a coarse account of classes. And Linnebo would after all be vindicated in thinking that the problem of identity for Maddy's

[12]Weak entailment and strong entailment in Field's system correspond to global and local consequence in a supervaluationist setting. For more on the relationship between global and local consequence, where the latter is characterized using the material conditional, see Incurvati and Schlöder, 2022.

theory should lead us to conclude that a hierarchical approach to classes based on a Kripke-style construction is a failed research programme.

8. Maximal consistency

Fortunately, we can give a coarse account of classes while remaining within the remit of a hierarchical approach to classes based on a Kripke-style construction. To develop such an account, a proof-theoretic point of view will again prove useful. The fact that ($+\simeq$ I.) disallows the use of the η-rules in its subderivations sheds light on why we cannot conclude that $\hat{x}(x \not\eta x)$ and $\hat{x}(x \not\eta x \wedge x \not\eta x)$ are identical. For it is rather natural to try to establish that $\hat{x}(x \not\eta x)$ and $\hat{x}(x \not\eta x \wedge x \not\eta x)$ via the following derivation:

$$\dfrac{\dfrac{\dfrac{\dfrac{[+t\eta\hat{x}(x \not\eta x)]}{+t \not\eta t}(+\eta \text{ E.}) \quad \dfrac{[+t\eta\hat{x}(x \not\eta x)]}{+t \not\eta t}(+\eta \text{ E.})}{+t \not\eta t \wedge t \not\eta t}(+\wedge \text{ I.})}{+t\eta\hat{x}(x \not\eta x \wedge x \not\eta x)}(+\eta \text{ I.}) \quad \dfrac{\dfrac{\dfrac{[+t\eta\hat{x}(x \not\eta x \wedge x \not\eta x)]}{+t \not\eta t \wedge t \not\eta t}(+\eta \text{ E.})}{+t \not\eta t}(+\wedge \text{ E.})}{+t\eta\hat{x}(x \not\eta x)}(+\eta \text{ I.})}{+t\eta\hat{x}(x \not\eta x) \simeq t\eta\hat{x}(x \not\eta x \wedge x \not\eta x)}(+\simeq \text{ I.})$$

However, this derivation is not valid in MvFBL because it makes use of the η-rules in the subderivations of the $\simeq$ introduction rule. Now, the ban on applications of η-rules in subderivations was motivated on the grounds that, from a supervaluationist standpoint, classical reasoning is in order when no rules are used that can engender indeterminacy. But from a supervaluationist standpoint, classical reasoning would also seem to be in order when, even though the rules are used, they are used only to access and therefore exploit the logical structure of the defining formula F of some class $\hat{x}F$. For this structure and what follows from it are unaffected by the way in which the indeterminacy might be resolved. We can formally capture this idea by weakening the restriction on the $\simeq$ introduction rule so that the use of the η-rules is allowed when the subderivations proceed first by eliminating η and then introducing it at the end.

$$(+\simeq \text{ I.}) \dfrac{\begin{array}{cc}[+t\eta\hat{x}A] & [+t\eta\hat{x}B] \\ \vdots & \vdots \\ +t\eta\hat{x}B & +t\eta\hat{x}A\end{array}}{+\hat{x}A \simeq \hat{x}B}$$

if either the subderivations use no η-rules or they begin with an application of ($+\eta$E.) and ($-\eta$E.) to every premiss and end with an application of ($+\eta$I.) or ($-\eta$I.)

With this weaker restriction in place, the above derivation of $+t\eta\hat{x}(x \not\eta x) \simeq t\eta\hat{x}(x \not\eta x \wedge x \not\eta x)$ goes through, as it perhaps should.

Clearly, however, if the motivation for weakening the restriction applies to the introduction rule for asserted class identity, it ought to apply whenever in MvFBL we had restrictions on the subderivations. Indeed, there is also a technical reason for this, namely that we are officially treating the introduction rule for asserted class identity as a derived rule, and in order to be able to derive the more permissive version of it, we must relax the restrictions on the subderivations on the (BET*) rule accordingly. This yields the following rule.

$$\text{(BET**)}\ \dfrac{\begin{array}{cc}[+A] & [-A] \\ \vdots & \vdots \\ \varphi & \varphi\end{array}}{\varphi}$$

if either the subderivations use no η-rules or they begin with an application of ($+\eta$E.) and ($-\eta$E.) to every premiss and end with an application of ($+\eta$I.) or ($-\eta$I.)

A similar change must be made in the restrictions on the subderivations in the ($-\wedge$E.) and ($-\forall$E.) rules. Call (for reasons that will shortly become clear) the resulting theory MCBL. If we let a $\clubsuit_i$ denote one of the two force markers, MCBL can also be axiomatized by adding to MvFBL the following meta-rule of Classical Compositionality, which perhaps more directly captures the idea that classical reasoning is in order when the η-rules are only used to access the logical structure of the defining condition of the relevant class.[13]

$$\text{(CC)}\ \dfrac{\clubsuit_1 F_1 t_1, \ldots, \clubsuit_n F_n t_n \vdash_{\mathsf{MvFBL}^*} \clubsuit_o G u}{\clubsuit_1 t_1 \eta \hat{x} F_1, \ldots, \clubsuit_n t_n \eta \hat{x} F_n \vdash_{\mathsf{MvFBL}^*} \clubsuit_o u \eta \hat{x} G}$$

We have arrived at MCBL via proof-theoretic considerations and by reflecting on the original motivation for the restrictions on hypothetical reasoning within a supervaluationist setting. What is remarkable is that MCBL is also a very natural theory from a model-theoretic point of view.

To explain and make precise the sense in which this is the case, we need to introduce some further model-theoretic terminology. I mentioned earlier that Kripke (1975) had already suggested the possibility of using a supervaluational scheme for handling truth-value gaps, and that the first scheme he considered was a straightforward adaption of van Fraassen's original notion of supervaluation to the case of truth. But in his paper, Kripke also considered another supervaluational scheme. The idea of this scheme is to take an expansion of an EA-pair to be admissible just in case it is maximally consistent. This gives rise to the following definition.

Definition 8.1 (mc-admissible expansion)**.** An EA-pair $\tau = \langle \tau^+, \tau^- \rangle$ is a *mc-admissible expansion* of an EA-pair $\sigma = \langle \sigma^+, \sigma^- \rangle$ iff (i) $\tau \sqsupseteq \sigma$, (ii) for no A, both $\tau \models A$ and $\tau \models\!\!| A$, and (iiii) for every A, either $\tau \models A$ or $\tau \models\!\!| A$.

Using this notion in the Supervaluational-jump template (Definition 6.1), we then obtain the maximally consistent jump J_{mc}, and replacing the Maddy jump with J_{mc} in the definition of the Maddy hierarchy, we obtain the *maximally consistent class hierarchy*. Once again, the construction reaches a fixed point σ^{mc}, and so we can consider σ^{mc} as providing the extension and the anti-extension of η according to the theory of classes obtained by adopting the maximally consistent supervaluational scheme within the context of a hierarchical theory of classes.

We are now ready to prove that there is a deep connection between MCBL and the maximally consistent class hierarchy. For, as I am now going to show, over TST, MCBL axiomatizes the fixed point of this hierarchy.[14] I first prove that MCBL is sound with respect to the consequence relation induced by σ^{mc}.

[13] The name is intended to be reminiscent of compositionality in the truth context, where two special cases of this rule have been shown by Incurvati and Schlöder (2023) to deliver material compositionality for the truth predicate. Indeed, the situation suggests to me that the problem of identity in the theory of classes is closely linked to the problem of compositionality in the theory of truth. I hope to explore these connections in future work.

[14] The result mirrors Incurvati and Schlöder's (2023) result that a suitable extension of their multilateral theory of truth axiomatizes the maximally consistent truth hierarchy.

Theorem 8.2. *Let Γ be a set of signed sentences and φ a signed sentence. Suppose that for all ψ in Γ, $\sigma^{\mathrm{mc}} \models A$ if $\psi = +A$, and $\sigma^{\mathrm{mc}} \not\models A$ if $\psi = -A$. Then, if $\Gamma \vdash_{\mathsf{MCBL}} \varphi$, $\sigma^{\mathrm{mc}} \models B$ if $\varphi = +B$ and $\sigma^{\mathrm{mc}} \not\models B$ if $\varphi = -B$.*

Proof. Recall that MCBL can be axiomatized by adding the Classical Compositionality metarule to MvFBL. Now the arguments from 6.3 carry over to the present case, except that for the soundness of ($+\eta$E.) we appeal to the fixed-point property. So it remains to check the soundness of (CC).

Let $\circ$ denote nothing if a given $\clubsuit_i$ is $+$ and denote negation if a given $\clubsuit_i$ is $-$. Then since $\vdash_{\mathsf{MvFBL}^*}$ denotes derivability in MvFBL without use of the η-rules, to prove that (CC) is sound, it suffices to show that if $\sigma^{\mathrm{mc}} \models \circ F_1 t_1 \wedge \ldots \wedge \circ F_n t_n \wedge \circ \neg G u$, then $\sigma^{\mathrm{mc}} \models \circ t_1 \eta \hat{x} F_1 \wedge \ldots \wedge \circ t_n \eta \hat{x} F_n \wedge \circ u \not\eta \hat{x} G$. So suppose that $\models \circ F_1 t_1 \wedge \ldots \wedge \circ F_n t_n \wedge \circ \neg G u$ and, for *reductio*, that $\sigma^{\mathrm{mc}} \not\models \circ t_1 \eta \hat{x} F_1 \wedge \ldots \wedge \circ t_n \eta \hat{x} F_n \wedge \circ u \not\eta \hat{x} G$. This means that there is an mc-admissible expansion τ of σ^{mc} such that $\tau \not\models \circ t_1 \eta \hat{x} F_1 \wedge \ldots \wedge \circ t_n \eta \hat{x} F_n \wedge \circ u \not\eta \hat{x} G$. It follows that $\tau \not\models \circ F_1 t_1 \wedge \ldots \wedge \circ F_n t_n \wedge \circ \neg G u$. But since τ is maximally consistent, this means that $\tau \not\models \circ F_1 t_1 \wedge \ldots \wedge \circ F_n t_n \wedge \circ \neg G u$. But since τ is an expansion of σ^{mc}, we also have that $\tau \models \circ F_1 t_1 \wedge \ldots \wedge \circ F_n t_n \wedge \circ \neg G u$. Hence, τ is not classically consistent and hence not an mc-admissible expansion. □

Similarly to the cases of MBL and MvFBL, Theorem 8.2 yields that, over TST, MCBL is sound with respect to σ^{mc}.

Theorem 8.3. *For every sentence A, if $\mathsf{TST} \vdash_{\mathsf{MCBL}} +A$, then $\sigma^{\mathrm{mc}} \models A$, and if $\mathsf{TST} \vdash_{\mathsf{MCBL}} -A$, then $\sigma^{\mathrm{mc}} \not\models A$.*

Given our earlier results and proofs, it is then easy to see that, over TST, MCBL is also complete with respect to σ^{mc}.

Theorem 8.4. *For every sentence A, if $\sigma^{\mathrm{mc}} \models A$, then $\mathsf{TST} \vdash_{\mathsf{MCBL}} +A$, and if $\sigma^{\mathrm{mc}} \not\models A$, then $\mathsf{TST} \vdash_{\mathsf{MCBL}} -A$.*

Proof. The proofs of Lemma 6.5 and Theorem 6.6 go through as before. The same goes for the proof of Theorem 6.7, except that we now need to show that the τ defined in the course of the proof, besides satisfying condition (i) of the definition of an mc-admissible expansion of $\sigma^{\mathrm{mc}}_{\alpha}$ (that is, being an expansion of it), also satisfies conditions (ii) and and (iii) (that is, being consistent and maximal).

For consistency, suppose that $\tau \models A$ and $\tau \not\models A$ for some A. Then, by Theorem 6.6 adapted to the case of MCBL, it follows that $\Gamma \vdash_{\mathsf{MCBL}^*} +A$ and $\Gamma \vdash_{\mathsf{MCBL}^*} -A$, contradicting the assumption that Γ is b-consistent. The maximality of τ can be established in a similar manner. □

It is then straightforward to provide a general model theory for MCBL by requiring the models to be not only η-admissible but, in effect, to satisfy Classical Compositionality.

9. Conclusion

Far from being a failed research programme, the Maddian approach to classes is alive and well, once we pursue it along the supervaluational lines already mentioned by Kripke in 1975. The rampant indeterminacy of Maddy's original approach is not inherent to a hierarchical approach to classes based on a Kripke-style construction. Rather, it is due to the particular implementation using the scheme for handling truth-value gaps Kripke focused on, which gives rise to a Strong

Kleene Logic. By using a different, supervaluational scheme, we can address the basic problem of rampant indeterminacy. And by focusing on the specific supervaluational scheme based on maximally consistent expansions, we can give a natural solution to the longstanding problem of identity for a naïve theory of classes, meeting a more general challenge for naïve theories of classes issued by Restall. Looking at the problem from a proof-theoretic perspective, and providing bilateral and multilateral deductive systems for the theories of classes examined, played a key role in arriving at this solution, in that it made it clear exactly what shape a rule of identity among classes ought to have.

Many questions remain open. In closing, let me highlight two sets of questions that I deem especially pressing. The first is that of the conditional in MCML. Although the theory includes the unrestricted η-rules, it does not validate the comprehension schema where the schema is formulated using the material conditional of the theory. According to Field et al. (2017), any naïve theory of classes worth its name ought to validate the unrestricted comprehension schema. Field et al. (2017) go on to prove an impossibility result concerning extensionality in the setting of a naïve theory of classes: in the presence of what they deem to be very modest demands on extensionality and the conditional, it is not possible to consistently validate the unrestricted comprehension schema.

Now, Incurvati and Schlöder (2023, Ch. 8) have developed a theory of conditionals on which one of the principles Field et al. (2017) use in their proof (namely, Quasi-Substitutivity) appears to fail. For this reason, it might be profitable to investigate whether it is possible to validate a version of the comprehension schema formulated using the Incurvati-Schlöder indicative conditional:

Question 9.1. Is the theory obtained by adding the comprehension schema formulated using the Incurvati-Schlöder conditional to MCML consistent?

If the answer to this question is positive, a subsequent question would then be whether the Incurvati-Schlöder conditional suffices to sustain a decent amount of conditional reasoning within the theory.

This brings me to the second set of questions, which concern the mathematical strength of the maximally consistent theory of classes. Much is not known here at the time of writing. For instance, it is not known whether a fully satisfactory theory of Frege–Russell cardinal numbers can be developed within MCML. We know that MCML improves on Maddy's theory of classes by establishing basic cardinality facts. For instance, while Maddy's theory could not even prove that $\{\emptyset\}$ is not equinumerous with $\{\emptyset, \{\emptyset\}\}$, this holds in the maximally consistent theory of classes. The intuitive reason for this is that in any model over which we are supervaluating there is no class witnessing the one-to-one correspondence between $\{\emptyset\}$ and $\{\emptyset, \{\emptyset\}\}$, and so no such correspondence exists in the universe of the maximally consistent theory of classes. However, it seems plausible to require that MCML should be able to prove Hume's Principle, formulated as $\hat{x}(x \approx t) = \hat{x}(x \approx t') \leftrightarrow t \approx t'$.

Question 9.2. Is Hume's Principle provable in MCML?

Given the difficulties surrounding the development of a theory of Frege–Russell cardinal numbers within her theory of classes, Maddy (2000, 313–314) explores the prospects of developing a theory of cardinal numbers based on von Neumann ordinals. While the results are more encouraging in this case, the theory still falls short of deriving arithmetic. In particular, the

Induction Axiom is not provable because of the problem of rampant indeterminacy. One might therefore ask:

Question 9.3. Are the Peano Axioms provable in MCML given a definition of cardinality *à la* von Neumann?

In the theory of truth, much work has been devoted to establishing the proof-theoretic strength of various axiomatizations of the truth predicate. For instance, and relevantly for our purposes, it was established by Andrea Cantini (1990) that a theory of truth using a more demanding supervaluational scheme than the van Fraassen one but less demanding than the maximally consistent one has the same proof-theoretic strength as the theory of inductive definitions known as ID_1. For this reason, it is to be expected that the theory of classes developed here will have considerable proof-theoretic strength. I therefore ask:

Question 9.4. What is the proof-theoretic strength of MCML?

Acknowledgements. This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 101086295) within the project *Philosophical, Logical, and Experimental Routes to Substructurality*. For comments and discussion, I would like to thank Maria Beatrice Buonaguidi, Orestis Dimou Belegratis, Pablo Dopico Fernandez, Øystein Linnebo, Penelope Maddy, Simone Picenni, Graham Priest, Lorenzo Rossi, and a referee for this journal. Earlier versions of this material were presented at the Graduate Center of the City University of New York, at the University of Turin, at the University of Pisa, and at the University of Calcutta. I am grateful to the members of these audiences for their valuable feedback.

References

Aczel, P. (1980). Frege structures and the notions of proposition, truth and set. In Barwise, J., Keisler, J., and Kunen, K., editors, *The Kleene Symposium*, volume 101 of *Studies in Logic and the Foundations of Mathematics*, pages 31–59, North Holland, Amsterdam. Elsevier. DOI: https://doi.org/10.1016/S0049-237X(08)71252-7.

Barton, N. and Williams, K. J. (2024). Varieties of class-theoretic potentialism. *The Review of Symbolic Logic*, 17:272–304. DOI: https://doi.org/10.1017/S1755020323000126.

Bealer, G. (1982). *Quality and Concept*. Clarendon Press, Oxford. DOI: https://doi.org/10.1093/acprof:oso/9780198244288.001.0001.

Boolos, G. (1974). Reply to Charles Parsons' "Sets and classes". In Boolos, 1998, 30–36.

Boolos, G. (1985). Nominalist platonism. *Philosophical Review*, 94:327–344. Reprinted in Boolos, 1998, 73–87. DOI: https://doi.org/10.2307/2185003.

Boolos, G. (1998). *Logic, Logic, and Logic*. Harvard University Press, Cambridge, Massachusetts.

Button, T. and Trueman, R. (2024). A fictionalist theory of universals. In Fritz, P. and Jones, N. K., editors, *Higher-Order Metaphysics*, pages 245–290, Oxford. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780192894885.003.0007.

Cantini, A. (1990). A theory of formal truth arithmetically equivalent to ID_1. *Journal of Symbolic Logic*, pages 244–59. DOI: https://doi.org/10.2307/2274965.

Chierchia, G. and Turner, R. (1988). Semantics and property theory. *Linguistics and Philosophy*, 11:261–302. DOI: https://doi.org/10.1007/BF00632905.

del Valle-Inclan, P. (2023). Harmony and normalisation in bilateral logic. *Bulletin of the Section of Logic*, 52:377–409. DOI: https://doi.org/10.18778/0138-0680.2023.14.

Ferrari, F. and Incurvati, L. (2022). The varieties of agnosticism. *The Philosophical Quarterly*, 72:365–380. DOI: https://doi.org/10.1093/pq/pqab038.

Field, H. (2008). *Saving Truth from Paradox*. Oxford University Press, New York. DOI: https://doi.org/10.1093/acprof:oso/9780199230747.001.0001.

Field, H. (2010). Replies to commentators on *Saving Truth from Paradox*. *Philosophical studies*, 147:457–470.

Field, H., Lederman, H., and Øgaard, T. F. (2017). Prospects for a naive theory of classes. *Notre Dame Journal of Formal Logic*, 58:461–506. DOI: https://doi.org/10.1215/00294527-2017-0010.

Fine, K. (1975). Vagueness, truth, and logic. *Synthese*, 30:265–300. DOI: https://doi.org/10.1007/BF00485047.

Florio, S. (2014). Unrestricted quantification. *Philosophy Compass*, 9:441–454. DOI: https://doi.org/10.1111/phc3.12127.

Frege, G. (1884). *Die Grundlagen der Arithmetik*. Wilhelm Koebner, Breslau. Translated as Frege, 1953.

Frege, G. (1892). On concept and object. *Vierteljahrsschrift für wisseschaftliche Philosophie*, 16:192–205. Translated in Geach and Black, 1952, 42–55.

Frege, G. (1953). *The Foundations of Arithmetic*. Basil Blackwell, Oxford, 2nd edition.

Fujimoto, K. (2019). Predicativism about classes. *The Journal of Philosophy*, 116:206–229. DOI: https://doi.org/10.5840/jphil2019116413.

Geach, P. and Black, M., editors (1952). *Translations from the Philosophical Writings of Gottlob Frege*. Basic Blackwell, Oxford.

Gilmore, P. C. (1974). The consistency of partial set theory without extensionality. In Jech, T. J., editor, *Axiomatic Set Theory II. Proceedings of Symposia in Pure Mathematics, 13*, pages 147–153, Providence, RI. American Mathematical Society.

Gitman, V., Hamkins, J. D., Holy, P., Schlicht, P., and Williams, K. J. (2020). The exact strength of the class forcing theorem. *The Journal of Symbolic Logic*, 85:869–905. DOI: https://doi.org/10.1017/jsl.2019.89.

Grišin, V. N. (1982). Predicate and set-theoretic calculi based on logic without contractions. *Mathematics of the USSR-Izvestiya*, 18:41. DOI: https://doi.org/10.1070/IM1982v018n01ABEH001382.

Hamkins, J. D., Kirmayer, G., and Perlmutter, N. L. (2012). Generalizations of the kunen inconsistency. *Annals of Pure and Applied Logic*, 163:1872–1890. DOI: https://doi.org/10.1016/j.apal.2012.06.001.

Hinnion, R. and Libert, T. (2003). Positive abstraction and extensionality. *The Journal of Symbolic Logic*, 68:828–836. DOI: https://doi.org/10.2178/jsl/1058448441.

Holy, P., Krapf, R., Lücke, P., Njegomir, A., and Schlicht, P. (2016). Class forcing, the forcing theorem and Boolean completions. *The Journal of Symbolic Logic*, 81:1500–1530. DOI: https://doi.org/10.1017/jsl.2016.4.

Horsten, L. and Linnebo, Ø. (2016). Term models for abstraction principles. *Journal of Philosophical Logic*, 45:1–23. DOI: https://doi.org/10.1007/s10992-015-9344-z.

Incurvati, L. (2020). *Conceptions of Set and the Foundations of Mathematics*. Cambridge University Press, Cambridge. DOI: https://doi.org/10.1017/9781108596961.

Incurvati, L. and Schlöder, J. J. (2017). Weak rejection. *Australasian Journal of Philosophy*, 95:741–60. DOI: https://doi.org/10.1080/00048402.2016.1277771.

Incurvati, L. and Schlöder, J. J. (2019). Weak assertion. *Philosophical Quarterly*, 69:741–770. DOI: https://doi.org/10.1093/pq/pqz016.
Incurvati, L. and Schlöder, J. J. (2022). Meta-inferences and supervaluationism. *Journal of Philosophical Logic*, 51:1549–1582. DOI: https://doi.org/10.1007/s10992-021-09618-4.
Incurvati, L. and Schlöder, J. J. (2023). Inferential deflationism. *Philosophical Review*, 132:529–578. DOI: https://doi.org/10.1215/00318108-10697531.
Incurvati, L. and Schlöder, J. J. (2023). *Reasoning With Attitude: Foundations and Applications of Inferential Expressivism*. Oxford University Press, New York. DOI: https://doi.org/10.1093/oso/9780197620984.001.0001.
Jones, N. K. (2016). A higher-order solution to the problem of the concept *horse*. *Ergo*, 3. DOI: https://doi.org/10.3998/ergo.12405314.0003.006.
Keefe, R. (2000). *Theories of Vagueness*. Cambridge University Press, Cambridge.
Kreisel, G. (1967). Informal rigour and completeness proofs. In Lakatos, I., editor, *Problems in the Philosophy of Mathematics*, pages 138–171, Amsterdam. North-Holland.
Kriener, J. (2014). The groundedness approach to class theory. *Inquiry*, 57:244–273. DOI: https://doi.org/10.1080/0020174X.2013.855657.
Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy*, 72:690–716. DOI: https://doi.org/10.2307/2024634.
Lear, J. (1977). Sets and semantics. *Journal of Philosophy*, 74:86–102. DOI: https://doi.org/10.2307/2025573.
Linnebo, Ø. (2006). Sets, properties, and unrestricted quantification. In Rayo, A. and Uzquiano, G., editors, *Absolute Generality*, pages 149–178. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/oso/9780199276424.003.0006.
Linnebo, Ø. (2024). Maddy on classes. In Arbeiter, S. and Kennedy, J., editors, *The Philosophy of Penelope Maddy*, pages 65–80. Springer, Switzerland. DOI: https://doi.org/10.1007/978-3-031-58425-1_6.
Maddy, P. (1983). Proper classes. *Journal of Symbolic Logic*, 48:113–139. DOI: https://doi.org/10.2307/2273327.
Maddy, P. (1990). *Realism in Mathematics*. Clarendon Press, Oxford. DOI: https://doi.org/10.1093/019824035X.001.0001.
Maddy, P. (2000). A theory of sets and classes. In Sher, G. and Tieszen, R., editors, *Between Logic and Intuition. Essays in Honor of Charles Parsons*, pages 299–316, Cambridge. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511570681.016.
Maddy, P. (2024). Reply to Linnebo. In Arbeiter, S. and Kennedy, J., editors, *The Philosophy of Penelope Maddy*, pages 81–84. Springer, Switzerland. DOI: https://doi.org/10.1007/978-3-031-58425-1_7.
Meadows, T. (2015). Infinitary tableau for semantic truth. *Review of Symbolic Logic*, 8:207–35. DOI: https://doi.org/10.1017/S175502031500012X.
Parsons, C. (1974). Sets and classes. *Noûs*, 8:1–12. DOI: https://doi.org/10.2307/2214641.
Quine, W. V. (1951). Two dogmas of empiricism. *Philosophical Review*, 60:20–43. DOI: https://doi.org/10.2307/2181906.
Quine, W. V. O. (1948). On what there is. *The Review of Metaphysics*, 2:21–38.
Rayo, A. and Uzquiano, G. (1999). Toward a theory of second-order consequence. *Notre Dame Journal of Formal Logic*, 40:315–325. DOI: https://doi.org/10.1305/ndjfl/1022615612.
Rayo, A. and Williamson, T. (2003). A completeness theorem for unrestricted first-order languages. In Beall, J. C., editor, *Liar and Heaps: New Essays on Paradox*, pages 331–356, Oxford. Clarendon Press. DOI: https://doi.org/10.1093/oso/9780199264803.003.0016.

Restall, G. (2010). What are we to accept, and what are we to reject, while saving truth from paradox? *Philosophical Studies*, 147:433–443. DOI: https://doi.org/10.1007/s11098-009-9468-5.

Rumfitt, I. (2000). "Yes" and "no". *Mind*, 109:781–823. DOI: https://doi.org/10.1093/mind/109.436.781.

Russell, B. (1903). *Principles of Mathematics*. Allen & Unwin, London.

Schindler, T. (2019). Classes, why and how. *Philosophical Studies*, 176:407–435. DOI: https://doi.org/10.1007/s11098-017-1022-2.

Simonelli, R. (Forthcoming). "Yes", "no", neither, and both. *Synthese*.

Smiley, T. (1996). Rejection. *Analysis*, 56:1–9. DOI: https://doi.org/10.1111/j.0003-2638.1996.00001.x.

Standefer, S., French, R., and Restall, G. (2020). Proofs and models in naive property theory: A response to Hartry Field's 'Properties, propositions and conditionals'. *Australasian Philosophical Review*, 4:162–177. DOI: https://doi.org/10.1080/24740500.2021.1886690.

Trueman, R. (2021). *Properties and Propositions: The Metaphysics of Higher-Order Logic*. Cambridge University Press, Cambridge. DOI: https://doi.org/10.1017/9781108886123.

van Fraassen, B. C. (1966). Singular terms, truth-value gaps, and free logic. *The Journal of Philosophy*, 63:481–495. DOI: https://doi.org/10.2307/2024549.

White, R. B. (1979). The consistency of the axiom of comprehension in the infinite-valued predicate logic of Łukasiewicz. *Journal of Philosophical Logic*, 8:509–534. DOI: https://doi.org/10.1007/BF00258447.

Journal for the Philosophy of Mathematics

OPEN ACCESS

Citation: LEITGEB, Hannes (2025). Carnapian Logicism and Semantic Analyticity. *Journal for the Philosophy of Mathematics*. 2: 75-106. doi: 10.36253/jpm-3468

Received: April 20, 2025

Accepted: November 1, 2025

Published: December 30, 2025

ORCiD
HL: 0000-0002-8276-149X

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Firenze University Press
www.fupress.com

Carnapian Logicism and Semantic Analyticity

HANNES LEITGEB

Faculty of Philosophy, Philosophy of Science and Religious Studies, Ludwig-Maximilians-University Munich, Germany.
Email: hannes.leitgeb@lmu.de

Abstract: This article argues for a (quasi-)Carnapian version of logicism about mathematics: there is a logicist conceptual framework in which (i) all standard mathematical terms are defined by logical terms, and (ii) all standard mathematical theorems are (likely to be) analytic. Along the way, the article explains the historical-philosophical background, how the definitions in (i) are to proceed, what the framework and the semantic notion of analyticity-in-a-framework are like, and why the probabilistic qualification 'likely to be' is used in (ii). The upshot is not some logicist epistemic foundationalism about mathematics but the insight that mathematics can be rationally reconstructed as being conceptual, i.e., as coming along with a conceptual framework.

Keywords: Logicism, mathematics, Carnap, analytic, probability.

1. Introduction

In the famous Königsberg conference from 1930, in which Arend Heyting presented intuitionism and John von Neumann formalism, Rudolf Carnap gave a lecture on logicism about mathematics which appeared as Carnap (1931) later. In his lecture, Carnap stated the logicist thesis in the following, still fairly standard, two-part manner:[1]

> 1. The *concepts* of mathematics can be derived from logical concepts through explicit definitions.
> 2. The *theorems* of mathematics can be derived from logical axioms through purely logical deduction. (Carnap, 1931, pp. 91f)

In line with Carnap, I am going to understand by *traditional logicism* about a mathematical language $\mathcal{L}$ and a mathematical theory $T_{\mathcal{L}}$ (formulated in $\mathcal{L}$) the conjunction of the following two theses:

[1]Some more recent (neo-)logicists, such as Hale and Wright (2001), might expand 'logical concepts' to 'logical or abstraction concepts' in 1, and some logicists might replace 2 by 'The *truths* of mathematics are logical truths (or analytic)'. For a survey of logicism, see Tennant (2013).

1a. All mathematical terms in $\mathcal{L}$ are explicitly definable from logical terms.
2a. All mathematical theorems in $T_{\mathcal{L}}$ are logically derivable from logical axioms and explicit definitions (the definitions claimed to exist by 1a).

Both Frege's (1884; 1893/18931903) logicism about arithmetic and Whitehead and Russell's (1910-1910 1913) logicism about general mathematics may be understood as aiming at traditional logicism in that sense, with suitable choices of $\mathcal{L}$ and $T_{\mathcal{L}}$.

Moreover, if one follows Frege (e.g. §3 of his *Grundlagen der Arithmetik*, 1884) in defining a sentence to be analytic just in case it is logically derivable from logical axioms and explicit definitions, one may reformulate 2a above in the equivalent manner:

2b. All mathematical theorems in $T_{\mathcal{L}}$ are (Frege-)analytic.[2]

Indeed, for any logicist whatsoever it is perfectly clear that definitions are just as indispensable for their own project as they are for mathematical practice itself. Since, according to our present-day understanding of logic, definitions are neither logical axioms nor logical rules, the logicist goal is thus not to show that mathematics is purely logical but rather that mathematics is analytic. When Carnap's thesis 2 above speaks of mathematical theorems being derivable from logical axioms through purely logical deduction, he simply understands 'logical' broadly enough to encompass also definitions.

Fast-forwarding more than thirty years to Carnap's autobiography in his Schilpp volume, Carnap describes his early encounter with Frege's logicism again in very similar terms:

> I had learned from Frege that all mathematical concepts can be defined on the basis of the concepts of logic and that the theorems of mathematics can be deduced from the principles of logic. Thus the truths of mathematics are analytic in the general sense of truth based on logic alone. (Carnap, 1963, p.46)

As the 'I had learned' suggests, Carnap retained his logicist convictions until that final stage of his career. But of course it is important not to overlook the 'can' here: for he had argued in his *Logical Syntax of Language* (Carnap, 1934, 1937) that mathematics could also be understood differently if reconstructed in an alternative framework (e.g. mathematical terms might instead be regarded as primitive non-logical terms; see Carnap, 1934, 1937, §84 and also §78). Carnap recommended to be tolerant about frameworks, as for him there was no fact of the matter which of them would be the "right" one to reconstruct mathematics within: there was a plurality of suitable formally precise frameworks available, and whether and how one preferred to reconstruct mathematics in any one of them reduced to the practical question of what choices would serve the specific aims of one's reconstruction to greater extent. However, it should also be clear that Carnap regarded the logicist reconstruction of mathematics in a logicist framework to be one of the options available, and he took it to be his preferred option for many salient purposes.[3]

[2] For a different understanding of Frege-analyticity, see Boghossian (1996).
[3] For more on Carnap's logicism, see Bohnert (1975) and, more recently, Marschall (2021). See Schiemer (2022) for a survey of logicism in logical empiricism more broadly.

In what follows, I will take up some of Carnap's ideas about theoretical terms, conceptual frameworks[4], and analyticity in order to develop a distinctively (quasi-) Carnapian version of logicism about mathematics. The ambition is not historical, but rather the goal is the systematic development and defense of a version of logicism on broadly Carnapian grounds.

The corresponding logicist thesis I want to argue for is:

There is a logicist conceptual framework, such that

1c. all standard mathematical terms are explicitly defined from logical terms in the framework,

2c. all standard mathematical theorems are likely to be (Carnap-)analytic, that is, semantically analytic, in the framework.

The term 'standard' in 1c and 2c is meant to apply to (reconstructions of) almost all terms and theorems of present-day pure mathematics. This will come about by rationally reconstructing, on logicist grounds, the language of second-order ZF set theory, which is known to allow for the definition of all standard mathematical terms used by pure mathematicians, and also the axiomatic system of second-order ZF set theory, which is known to allow for the derivation of all standard theorems proven by pure mathematicians so far. I will presuppose the deductive system of second-order logic (with Choice) to be genuinely logical and the second-order universal quantifier to range over all classes and class-relations of first-order individuals. The outcome will be a rational reconstruction[5] of pure set theory and indeed pure mathematics in the sense of clarifying, precisifying, systematizing, and interpreting set theory and mathematics from a logicist point of view while remaining close to set-theoretic and mathematical practice (though slight deviations from that practice are allowed for the sake of other virtues). The discussion of the logicist understanding of applied mathematics and of the role of mathematics in the empirical sciences and engineering will have to be left for a different occasion.[6]

Set theory is widely accepted by mathematicians to be one possible foundation for almost all of today's mathematics. That is: it is widely held that all standard mathematical terms are explicitly definable using only logical terms and the membership predicate $\in$, and that all standard mathematical theorems are derivable from first-order or second-order ZF with Choice and hence are true in all standard models of ZF with Choice or indeed true in *the* intended model (if there is just one). Call this common view 'ZFCism'.[7] What will Carnapian logicism add to this? ZFCism is a purely mathematical view that is not by itself a philosophical interpretation of mathematics and is compatible with different such interpretations. E.g., it might be given a special kind of realist interpretation of the following sort: metaphysically, sets and membership might be assumed to exist independently of reasoners and language, and set-theoretic truths might be assumed to be metaphysically necessary. Epistemologically, sets and membership

[4]For much of his career, Carnap would have spoken of constitution systems, languages or linguistic frameworks, though sometimes he also used the term 'conceptual framework', as in "many problems concerning conceptual frameworks seem to me to belong to the most important problems in philosophy" (Carnap, 1963, p.862). I prefer the term 'conceptual framework' in order to make clear that frameworks in that sense are not just syntactical but also involve semantic rules and semantic interpretation mappings. Conceptual frameworks correspond to the semantical systems or intensionally interpreted languages that became central to Carnap's work once he had taken his semantic turn, as exemplified by Carnap (1942, 1947/1956). In their most general form, conceptual frameworks encode syntactic, logical, semantic, pragmatic, epistemic, ontic, and other choices, and their construction, study, and application is philosophically useful (contra Maddy, 2007, Chapter 5) whenever philosophical concepts, theses or arguments depend on such choices. The logicism of this paper does depend on such choices.

[5]See Leitgeb and Carus (2024, Supp. D) for more on Carnap on rational reconstruction.

[6]But Carnap (1934, 1937, §84) rightly stresses that securing the applicability of mathematics to the empirical world is itself a vital part of the logicist project.

[7]I owe this terminology to an anonymous reviewer.

might be assumed to be epistemically accessible by quasi-perceptual means through which set-theoretic statements can be justified. Semantically, the membership predicate might be taken to be primitive,[8] to have its intended interpretation(s) in virtue of certain non-semantic facts, and the set-theoretic axioms might be regarded as synthetic. And so forth. Carnapian logicism will differ substantially from any such realist interpretation: it will stay closer to ZFCism itself, adding only a definition of sethood and membership in logical terms, assuming their interpretations to satisfy the set-theoretic axioms and to otherwise remain arbitrary, and regarding the set-theoretic axioms to be (likely to be) analytic. The resulting interpretation of set theory will be "thin" or "deflationary" in a sense similar to deflationary theories of truth in which the truth predicate is regarded as a (quasi-)logical expression the interpretation of which is only assumed to satisfy the Tarskian truth scheme for the object language in question. No substantial metaphysical or epistemological assumptions are assumed by any such deflationary conception of truth, and no such assumptions will be assumed by Carnapian logicism either. And the purpose of developing such a logicist interpretation in precise terms by rationally reconstructing mathematics in a logicist framework are strictly philosophical, not mathematical: the goal is not to give mathematicians a mathematically better foundation to work with—just as the realist view sketched before would not give mathematicians a mathematically better foundation—but rather to show that mathematics can be understood to be purely conceptual.

More generally, *(quasi-)Carnapian logicism* about a mathematical language $\mathcal{L}$ and a mathematical theory $T_{\mathcal{L}}$ is given by:

There is a logicist conceptual framework, such that

1d($\mathcal{L}$). all mathematical terms in $\mathcal{L}$ are explicitly defined from logical terms in the framework,

2d($T_{\mathcal{L}}$). all mathematical theorems in $T_{\mathcal{L}}$ are likely to be (Carnap-)analytic, that is, semantically analytic, in the framework.

Hence, 1c and 2c from before will follow from instantiating the schemes 1d and 2d with the names of the language $\mathcal{L}^2_{\in,Set}$ and the axiomatic system $ZF2[\in, Set]$ of second-order ZF. Focusing on these particular instances 1d($\mathcal{L}^2_{\in,Set}$) and 2d($ZF2[\in, Set]$) will prove useful for my logicist purposes; in particular, it will be convenient in so far as second-order set theory—just as its first-order variant—is regarded as at least one possible foundation for modern mathematics anyway. But most of the logicist project of the present paper could be carried out just as well for many other choices of $\mathcal{L}$ and $T_{\mathcal{L}}$. Therefore, readers are very much invited to apply the general logicist strategy of this article to other such choices, whether they concern alternative foundations of mathematics or, in a more piecemeal fashion, languages and theories for specific areas of mathematics (logicism about second-order Dedekind-Peano arithmetic, logicism about second-order Dedekind real analysis,. . .).

The definitions backing up 1d($\mathcal{L}^2_{\in,Set}$) will rely on an understanding of set-theoretic membership and sethood as theoretical concepts—concepts given by the axiomatic theory of second-order ZF—and on the corresponding explicit definitions of $\in$ and Set by purely logical higher-order epsilon terms (Section 2). The logicist framework in question, which is to be distinguished from proper scientific theories that can be formulated within the framework, will involve an object-linguistic and a metalinguistic part, both of which will

[8] In the usual axiomatic systems for set theory, '$\in$' is of course indeed a primitive. But this might change in a philosophical interpretation of set theory.

be based on higher-order logic and the logic of epsilon terms. The metalinguistic part will include semantic rules for the object-linguistic part, and a framework-relative semantic concept of analyticity will be introduced that is going to reflect these semantic rules (Section 3). Analyticity in that semantic sense will be entailed by Frege-analyticity, that is: logical axioms and explicit definitions in the object language of the framework and what is logically derivable from them in the framework will follow to be semantically analytic in the framework. But Carnapian semantic analyticity in the framework extends beyond Frege-analyticity in the framework. In other words: Frege-analyticity is sound but not complete with respect to semantic analyticity.[9] As we are going to see, the analyticity of second-order ZF in the semantic sense will depend on whether a certain second-order existence statement in the metalinguistic part of the framework holds true, which we will find very likely to be the case (Section 4). Accordingly, in contrast with traditional logicism, 2d($ZF2[\in, Set]$) does not claim the analyticity of standard mathematical theorems to be *derivable* from uncontroversial principles but just that these theorems are *likely to be* semantically analytic in the framework. Replacing the derivability of analyticity by its high probability should not come as too much of a surprise, as the analyticity of second-order ZF would entail its consistency, and we know from the Incompleteness Theorems that the consistency of second-order ZF could not be derived on more elementary grounds (assuming second-order ZF is consistent). Finally, I am going to draw some conclusions on what this (quasi-) Carnapian version of logicism does or does not show philosophically (Section 5). In particular, it would not serve any logicist version of epistemological foundationalism that would ask for logic to deliver a more secure foundation for mathematics. Instead, the main conclusions will be: mathematics can be rationally reconstructed as purely conceptual, that is, as coming along with a conceptual framework. In one version of Carnapian logicism, this includes the existence of abstract logical objects that are introduced by the logicist framework itself. And, at the very least, this reconstruction does not fare worse than any other philosophical interpretation of mathematics available, as it is formally clear, precise, and systematic, it remains close to mathematical practice, and it is philosophically coherent.[10]

Last but not least, I should stress that I have been qualifying my approach as *quasi*-Carnapian. The reason for this is that Carnap himself did *not* develop or defend logicism in this manner. Instead, for most of his logical and philosophical work, he relied on a version of the simple

[9] That is one reason why Carnapian logicism differs, e.g., from the conventionalism put forward by Warren (2020) who regards conventions as *syntactic* rules of language use. Another reason is that Warren's project is not one of rational reconstruction.

[10] The contemporary literature that comes closest to the theory to be presented are, first, Woods' (2014, Section 4.3) and Boccuni and Woods' (2020) version of abstractionist neo-logicism, second, Leitgeb, Nodelman, and Zalta's (2025) object-theoretic logicism, and third, Soysal's (2025) meta-semantic descriptivism. Woods and Boccuni advocate a neo-logicism in which abstraction operators in neo-logicist abstraction principles are given by second-order epsilon terms or, in any case, are semantically "arbitrary". The main differences to the present theory are: their philosophical background and interests consist in a combination of mathematical structuralism with Hale and Wright's (2001) neo-logicism based on abstraction principles (such as Hume's Principle); and they neither use a Carnapian concept of semantic analyticity nor argue for their basic principles to be analytic. Leitgeb, Nodelman, and Zalta (2025) also regard mathematical concepts as theoretical concepts given by mathematical theories. The differences are: they presuppose higher-order object theory as their background logic, which involves two kinds of predication; they do not invoke epsilon terms; and they do not apply a Carnapian concept of semantic analyticity. Instead, they combine Frege-analyticity with an extended notion of logical truth according to which a formula of a formal language is logically true just in case it is true in all models that include everything required for the possibility of having logically complex thoughts expressible in that language. As I will argue in Section 3, Carnapian frameworks are also meant to supply what is required for thought, and semantic analyticity in a framework tracks what the framework supplies. Finally, Soysal (2025) develops a version of meta-semantic descriptivism about logical and mathematical expressions in which these expressions have their meaning at least partially in virtue of descriptions; in the case of the membership predicate, the description is given by set-theoretic axioms. The differences from the present approach are: Soysal's theory is not a rational reconstruction but deals with the actual metasemantics of logical and mathematical language; it does not involve epsilon terms; and it is not meant to be based solely on Carnapian grounds (though overlapping with Carnap on theoretical terms and analyticity).

theory of types as his preferred logical system, which he described in formal detail in his *Abriß der Logistik* (Carnap, 1929), and from which parts of modern mathematics can be derived at least conditionally (that is, given certain assumptions, such as an Axiom of Infinity—see Carnap, 1929, Section 24e; I will return to this in Section 2). In fact, Marschall (2024, Section 3.2) presents historical reasons to believe that Carnap regarded our pre-theoretic understanding of set-theoretic membership to be sufficiently clear and determinate—much like Frege might have thought about the concept of extension—so that there would be no need to regard it as being determined by an axiomatic theory. But then again, as Bohnert (1975, p. 210) cites his conversation with Carnap in 1968, "He [Carnap] still thought set theory could be given an analytic interpretation".

In any case: what Carnap himself would have thought about the project of this paper is orthogonal to its strictly systematic ambitions. For me, the more interesting point is that Carnap *could* have thought of mathematics in the logicist manner I am going to describe, since he did have the philosophical resources to so so. And even more importantly: everyone else is invited to think of mathematics in the same manner, and if one did, one would be able to do so coherently.

2. Defining Membership and Sethood Logically

As mentioned in the introduction, the starting point of our considerations is the axiomatic theory

$$ZF2[\in, Set]$$

that is, classical second-order Zermelo-Fraenkel set theory (see Shapiro, 1991, p. 85)[11], which is formulated in the language $\mathcal{L}^2_{\in,Set}$, that is, with the logical and auxiliary symbols of classical second-order logic, the primitive descriptive binary membership predicate $\in$, and the primitive descriptive unary predicate Set for sets. Without further argument and Quinean worries notwithstanding, I will take the operators of pure second-order logic to be properly logical symbols, the axioms of the deductive system of second-order logic to be properly logical truths, and the rules of the deductive system of second-order logic to be properly logical valid.[12] The role of Set is just to restrict all first-order and second-order quantifiers in the axioms to sets, and to restrict the relata of the membership relation to sets.[13] In addition, I am also going to assume the Axiom of Extensionality for second-order entities to belong to the system of second-order logic; consequently, e.g., second-order property variables may be thought of as ranging over extensional properties or classes. And I will regard a second-order version of the Axiom of Choice to be included in the deductive system of second-order logic (following Shapiro, 1991, p. 67), without defending its logicality here.

[11] '$ZF2[\in, Set]$' can be used to denote the set of axioms of second-order set theory or the deductively closed set of formulas that are derivable from these axioms in the deductive system of second-order logic. The context should always make clear which of the two is meant in each case. In any case, I do not mean the set of formulas that are second-order consequences of these axioms in the model-theoretic sense. Similarly, I will leave it to the context to determine whether '$\in$' and 'Set' denote predicates, that is, linguistic items, or the concepts expressed by these predicates, or the extensions of these concepts.

[12] The deductive system of second-order logic may be viewed as a many-sorted variant of first-order logic and should thus be compatible even with a Quinean conception of logic.

[13] So all first-order quantifier occurrences in $ZF2[\in, Set]$ are of the form $\forall x(Set(x) \rightarrow \ldots)$ or $\exists x(Set(x) \wedge \ldots)$, the second-order Axiom of Replacement begins with $\forall f(\forall x(Set(x) \rightarrow Set(f(x))) \rightarrow \ldots)$, and the statement $\forall x, y(x \in y \rightarrow Set(x) \wedge Set(y))$ is accepted as yet another axiom. Note that overall $\forall x(Set(x) \leftrightarrow \exists y\ x \in y)$ becomes derivable. Given our aim of reconstructing pure mathematics, there will be no need to consider sets of urelements, that is, sets of non-sets. In a context in which no other entities were relevant than sets, the Set predicate could of course be eliminated, as is the case in Shapiro (1991, p. 85).

One of the advantages of going second-order is that second-order quantification makes the usual axiom schemes of first-order set theory obsolete, so that the axioms of $ZF2[\in, Set]$ may be regarded to form one longish but finite conjunction. But $ZF2[\in, Set]$ exhibits also other attractive features: almost all proven theorems of pure mathematics are known to be derivable from the axioms of $ZF2[\in, Set]$ and auxiliary definitions in the deductive system of second-order logic with Comprehension and Choice. Indeed, $ZF2[\in, Set]$ is a non-conservative extension of first-order ZFC, which, in turn, is often regarded as a foundation of modern pure mathematics. However, historically, Zermelo (1930) had formulated set theory (with urelements) in second-order terms, and second-order set theory seems to be closer to mathematical practice than its first-order version (see Shapiro, 1991, Sections 5.3–5.4). Like ZFC, $ZF2[\in, Set]$ captures the cumulative hierarchy of sets by proving that every set occurs in a hierarchy that is indexed by ordinals and given by $V_0 = \varnothing$, $V_{\alpha+1} = \wp(V_\alpha)$, and $V_\lambda = \cup_{\alpha<\lambda} V_\alpha$. In addition, unlike ZFC, $ZF2[\in, Set]$ is "almost" categorical, that is, it pins down the structure of the cumulative hierarchy uniquely up to its strongly inaccessible ordinal height (see Shapiro, 1991, p.86, and, for internal categoricity, Väänänen and Wang, 2015). This holds even though $ZF2[\in, Set]$ is of course deductively incomplete, as follows from the Incompleteness Theorems (assuming that $ZF2[\in, Set]$ is consistent). Moreover, for a logicist endeavor, it is reassuring that the basic individual entities described by $ZF2[\in, Set]$ are governed by the clear and precise identity criterion of first-order extensionality, they are similar in that way to Frege's extensions (which Frege regarded as logical objects), and they might even be viewed as intensionally rigid logical properties, so that e.g. a set $\{\varnothing, \ldots\}$ could be identified with the logical property $\lambda x(x = \varnothing \vee \ldots)$, the set $\varnothing$ with the logical property $\lambda x(x \neq x)$, and the like.[14]

Independently of whatever else the predicates $\in$ and Set might have meant antecedently, let us from now on think of $ZF2[\in, Set]$ as "implicitly defining" $\in$ and Set jointly with their underlying iterative conception of sets, where the iterative conception is preferably understood in a minimalist or deflationary manner (see Incurvati, 2020, Chapter 2 and especially Section 2.6). The fact that $ZF2[\in, Set]$ captures the cumulative hierarchy and is quasi-categorical goes some way towards making this plausible. Then what the theory $ZF2[\in, Set]$ does, next to its explicit or implicit existential claims about sets, is to determine the meanings of $\in$ and Set from their conceptual roles vis-a-vis the remaining meaningful symbols in $ZF2[\in, Set]$, that is, the logical symbols. And for these logical symbols I will take for granted that their meanings are fixed and determined uniquely. In particular, $\forall$ indeed means *for all*, whether on the first-order level of all individuals or on the second-order level of all classes, relations, and functions.[15] On that basis, the first step of our logicist reconstruction will consist in making the "implicit definition" of $\in$ and Set by $ZF2[\in, Set]$ fully explicit. In the remainder of this section, I am going to explain the idea of how this can be done, while the concrete implementation of that idea in a logicist framework will be carried out in the next section.

Now, what does it mean to understand $\in$ and Set so that all there is to them is given by $ZF2[\in, Set]$? Consider the open formula

$$ZF2[R, S]$$

[14] Carnap (1956, §23) discusses this option of reducing extensions to what he calls "L-determinate intensions". The option would not assume that all of these intensionally determinate logical properties could be expressed linguistically, of course.

[15] This is compatible with the deductive system of second-order logic being incomplete with respect to the "full" standard (model-theoretically defined) semantics for second-order logic. The deductive incompleteness of the system does not entail the expressive incompleteness of its language.

that results from replacing all occurrences of $\in$ in $ZF2[\in, Set]$ by the binary relation variable R and all occurrences of Set in $ZF2[\in, Set]$ by the unary class variable S. $ZF2[R, S]$ thereby expresses a constraint on the values of R and S. The idea will be to use $ZF2[R, S]$ to give a rigorous answer to the previous question by defining $\in$ and Set to be *an R and an S, respectively, such that $ZF2[R, S]$*. Other than satisfying the constraint expressed by $ZF2[R, S]$, the meanings of $\in$ and Set will be left arbitrary.

In more formal terms: let us assume our logical vocabulary to include Hilbert's indefinite description operator ϵ (see Hilbert and Bernays, 1934/19341939), both on the first-order and on the second-order level. Just as the standard definite description operator ι can be used to denote something by describing *the* entity that has such-and-such a property, ϵ can be used to denote something by describing *an* entity that has such-and-such a property. Thus, the epsilon operator is just like the iota operator but with the uniqueness presupposition stripped away. When there is more than one entity with the relevant property, the epsilon operator is instead understood to "pick" *any* of these entities. Accordingly, for first-order epsilon terms $\epsilon x\, \varphi[x]$ ("an x, such that $\varphi[x]$") and second-order epsilon terms $\epsilon R \psi[R]$ ("an R, such that $\psi[R]$"), the following two schemes comprise the logic of the epsilon operator (the so-called epsilon calculus, see Avigad and Zach, 2024):

Logical Axiom Scheme for First-Order Epsilon Terms:
$\vdash \exists x\, \varphi[x] \rightarrow \varphi[\epsilon x\, \varphi[x]]$.
Logical Axiom Scheme for Second-Order Epsilon Terms:
$\vdash \exists R \psi[R] \rightarrow \psi[\epsilon R\, \psi[R]]$.

When the antecedent of an instance of either of the schemes is false, so that there is no entity with the required property, no constraint is being imposed on what gets "picked" by the respective epsilon term.[16] With the exception of cases in which $\varphi[x]$ or $\psi[R]$ describes its respective x or R uniquely, I neither assume that there is a fact of matter of what is being "picked" by the respective epsilon term nor that there is a metasemantic mechanism that would determine what is being "picked". The idea is rather to view, e.g., the choice expressed by $\epsilon R \psi[R]$ as being describable in metalinguistic natural language terms by '$\epsilon R \psi[R]$ chooses *some/any/whatever* R, such that $\psi[R]$', which does not ascribe any fixed or determinate denotation to $\epsilon R \psi[R]$.[17]

While the Hilbert school used first-order epsilon terms in their efforts to carry out Hilbert's formalist program, Carnap (1959, [2000]) proposed to invoke second-order epsilon terms for the rational reconstruction of theoretical terms in science (see also Carnap, 1961). But before explaining Carnap's proposal in more detail, let me first state how second-order epsilon terms can be used to define $\in$ and Set explicitly.

I will present these definitions in two versions, the first of which is:

Definition 1. (Definition of $\in$ and Set, First Version)

(i) $\in =_{df} \epsilon R \exists S\, ZF2[R, S]$.
(ii) $\forall x$: $Set(x) \leftrightarrow_{df} \exists y\, x \in y$.

[16] One might also assume an extensionality axiom for ϵ to belong to the logic of ϵ, but it will not be relevant in what follows.

[17] If the metalanguage gets formalized itself, this amounts to interpreting object-linguistic epsilon terms by means of metalinguistic epsilon terms, which is much like the common practice of, e.g., stating the truth conditions of object-linguistic negation sentences with the help of the metalinguistic negation sign; see Section 3. See Leitgeb (2023) for a general semantic and metasemantic treatment of languages with semantically indeterminate expressions by means of metalinguistic epsilon terms.

Here, $\in$ is defined to be an(y) R, such that there is an S, such that $ZF2[R, S]$. And it is easy to see that, if there is such an R, the class Set is effectively defined to be the field of the relation $\in$ that has been defined by 1.[18] Thus, $\in$ and Set are defined by describing their conceptual roles in $ZF2[\in, Set]$, as promised.

The second version invokes a predicate '*Logical-in-*$\mathfrak{C}$' that is regarded as a primitive logical term by which logical individuals can be distinguished from non-logical ones in the conceptual framework $\mathfrak{C}$, where '$\mathfrak{C}$' denotes the forthcoming logicist framework:

Definition 2. (Definition of $\in$ and Set, Second Version)

(i) $\in =_{df} \epsilon R \exists S(\forall x(S(x) \rightarrow \textit{Logical-in-}\mathfrak{C}(x)) \wedge ZF2[R, S])$.
(ii) $\forall x$: $Set(x) \leftrightarrow_{df} \exists y\, x \in y$.

The basic idea is the same as before, it is just that $\in$ and Set are now stated explicitly to apply to logical objects only.

Mostly, I am going to focus on Definition 1, but for some purposes it will be useful to consider Definition 2 as an alternative, as will become clear in Section 4.

But why turn to epsilon terms at all and not use iota terms in these definitions? Indeed, following up Carnap's proposal, Lewis (1970) suggested to define theoretical terms in science by iota terms. And in the case of theories from empirical science, one might perhaps hope for these theories to describe the intended denotation of theoretical terms uniquely. However, in the case of a purely mathematical theory, any hope for uniqueness would be vain: for in all interesting cases, permutations of the underlying first-order domain would give rise to isomorphic but numerically distinct interpretations of the mathematical terms involved, which is why the uniqueness presupposition of iota terms would be violated. In contrast, non-uniqueness is unproblematic when mathematical concepts are defined by epsilon terms. What is more, the logicism I want to develop does not care about "the" intended interpretations of $\in$ and Set other than they satisfy $ZF2[R, S]$. Since $ZF2[R, S]$ is quasi-categorical (since $ZF2[\in, Set]$ is) and only includes logical expressions, one might also say: it only cares about the joint *logical structure* of $\in$ and Set (up to ordinal height). Structuralists about mathematics will concur,[19] though some non-structuralist realists about mathematics may not. But then again the task is to develop a logicist reconstruction of mathematics, not any such realist one.

There are further advantages to defining membership by a second-order epsilon term: assume that future set theorists will propose some new (say, large cardinal) axiom to be added coherently to $ZF2[\in, Set]$, and the mathematical community will go along with their proposal and regard the resulting system $ZF2^*[\in^*, Set^*]$ as their new foundation of mathematics. Then the "old" defining epsilon term $\epsilon R \exists S ZF2[R, S]$ from, say, Definition 1, could still be thought to denote the very same relation that a correspondingly updated logicism about $ZF2^*[\in^*, Set^*]$

[18] If one preferred, one could also define Set by yet another epsilon term, $Set =_{df} \epsilon X \forall x(X(x) \leftrightarrow \exists y\, x \in y)$, or, in this case (using second-order extensionality), even by $Set =_{df} \iota X \forall x(X(x) \leftrightarrow \exists y\, x \in y)$.

[19] See Boccuni and Woods (2020) for more on the affinity between (certain brands of) logicism and structuralism. See Leitgeb (2021) for the structuralist usage of epsilon terms to denote objects in ante rem structures: e.g., in an unlabeled graph G with two nodes and no edges, one may introduce a name a for one of these nodes by defining $a = \epsilon v(Vertex(v, G))$ and a name b for the other node by defining $b = \epsilon v(Vertex(v, G) \wedge v \neq a)$ (see Leitgeb, 2021, p. 79). a is then numerically distinct from b, but there is no non-semantic fact of the matter which of the two nodes is denoted by 'a' and which by 'b'. See Shapiro (2008, 2012) and Pettigrew (2008) for the closely related idea of regarding names for objects in structures, or for the structures themselves, as parameters introduced in the course of applying the logical rule of existential elimination. Schiemer and Gratzl (2016) also invoke epsilon terms in their reconstruction of structuralism.

would denote by its epsilon term $\epsilon R\,\exists S ZF2^*[R,S]$: although the equality

$$\epsilon R\,\exists S\; ZF2[R,S] = \epsilon R\,\exists S\; ZF2^*[R,S]$$

would not follow logically from $ZF2[\in, Set]$ and $ZF2^*[\in^*, Set^*]$, and even though there would be no fact of the matter whether that equality was true, it would be consistent with the two theories to accept it as true, and to speak as if membership as given by $ZF2[\in, Set]$ had always meant what would now be defined by means of $ZF2^*[\in^*, Set^*]$. It is that open-endedness of epsilon terms that Carnap aspired to exploit in his epsilon term reconstruction of theoretical terms, since he thought it nicely matched the open-endedness by which scientists may continue to specify the meanings of theoretical terms in the course of scientific development.[20] In the present context, it nicely matches the "inexhaustibility" of the concepts of set and membership that was described, e.g., by Gödel in his Gibbs lecture (Gödel, 1951 [1995]).[21]

Carnap's treatment of theoretical terms as epsilon terms may be viewed as a variant of his better known Ramsification reconstruction of a theoretical term T being given by a scientific theory $Th[T]$. He proposed to analyze $Th[T]$ in terms of what we now call the

Carnap sentence of $Th[T]$: $\exists R Th[R] \rightarrow Th[T]$

and the

Ramsey sentence of $Th[T]$: $\exists R Th[R]$.

The two of them taken together logically entail $Th[T]$ in second-order logic, and $Th[T]$ in turn logically entails their conjunction. Carnap's (1966) suggestion was to regard the Carnap sentence of $Th[T]$ as capturing the analytic content of $Th[T]$, since the only non-theoretical sentences (sentences without T) it entailed were logically true ones. In contradistinction, the Ramsey sentence of a typical theory $Th[T]$ from empirical science would capture the synthetic content of $Th[T]$, as it entailed the same non-theoretical (e.g. observation) sentences as $Th[T]$ itself.[22]

The correspondence to the epsilon term reconstruction of theoretical terms is: if T is defined by the epsilon term $\epsilon R Th[R]$, as suggested by Carnap (1959), the Carnap sentence of $Th[T]$ is indeed a logical consequence of that definition, which confirms its analytic status. Explained for the present context: since

(A) $\vdash \exists R \exists S\, ZF2[R,S] \rightarrow \exists S\, ZF2[\epsilon R\,\exists S\, ZF2[R,S], S]$

is an instance of the logical axiom scheme for second-order epsilon terms, 1 of Definition 1 combined with the Intersubstitutivity of Identicals yields the Frege-analyticity of

(B) $\exists R \exists S\, ZF2[R,S] \rightarrow \exists S\, ZF2[\in, S]$.

[20] "[. . .] this definition [by an epsilon term] gives just so much specification as we can give, and not more. We do not want to give more, because the meaning should be left unspecified in some respect, because otherwise the physicist could not—as he wants to—add tomorrow more and more postulates, and even more and more correspondence postulates, and thereby make the meaning of the same term more specific than it is today. So, it seems to me that the ϵ-operator is just exactly the tailor-made tool that we needed, in order to give an explicit definition, that, in spite of being explicit, does not determine the meaning completely, but just to that extent that it is needed" (Carnap, 1959, pp.171f).

[21] See Leitgeb (2023, Section 6) for more on the synchronic and diachronic advantages of dealing with semantic indeterminacy by means of epsilon terms.

[22] See Demopoulos (2007) and Suppl. E of Leitgeb and Carus (2024) for more on Carnap's reconstruction of theoretical terms in science.

And because $\forall x(S(x) \leftrightarrow \exists y\, x \in y)$ is logically derivable from (my formulation of) $ZF2[\in, S]$ in the deductive system of second-order logic, one can derive from (B) and 2 of Definition 1 the Carnap sentence

(C) $\exists R \exists S\, ZF2[R, S] \rightarrow ZF2[\in, Set]$,

which is thus Frege-analytic, too. Since Frege-analyticity will be seen to entail semantic analyticity in the next section, the Carnap sentence (C) of $ZF2[\in, Set]$ is therefore semantically analytic.

So far as the Ramsey sentence of $ZF2[\in, Set]$ is concerned, that is,

(R) $\exists R \exists S\, ZF2[R, S]$,

Sections 3 and 4 taken together will argue it to be likely to be semantically analytic, too, unlike the synthetic Ramsey sentences of typical empirical theories. And since semantic analyticity will be closed under logical derivability—and hence $ZF2[\in, Set]$ will be semantically analytic if (C) and (R) are—it will follow that $ZF2[\in, Set]$ is likely to be semantically analytic, just as promised. The same considerations apply *mutatis mutandis* to Definition 2 and its correspondingly expanded Carnap and Ramsey sentence.

But before I turn to the semantic analyticity of the Ramsey sentence (R), I will argue for the following: (i) Both Definition 1 and 2 indeed yield thesis 1d($\mathcal{L}^2_{\in,Set}$) from Section 1 for the language $\mathcal{L}^2_{\in,Set}$ of $ZF2[\in, Set]$. (ii) Even independently of the forthcoming argument for the analyticity of (R), Definition 1 (and analogously Definition 2) just by itself already amounts to a decent form of logicism, even when it does not quite deliver thesis 2d($ZF2[\in, Set]$) from Section 1.

About (i): given Definition 1, there are strong arguments in favor of 1d($\mathcal{L}^2_{\in,Set}$). For the only potentially non-logical terms in $\mathcal{L}^2_{\in,Set}$ are $\in$ and Set, and both of them are defined explicitly by purely logical terms: $\in$ is defined by the epsilon term $\epsilon R \exists S\, ZF2[R, S]$, which consists of only logical symbols, and since $\in$ is defined logically, the same holds for Set which is defined explicitly from $\in$. If there were a possible point of contention at all, it would concern whether the epsilon operator ϵ should count as logical. But see Woods (2014) for an argument to the effect that ϵ is logical in the Tarskian sense of permutation-invariance. Indeed, ϵ is very closely related to the existential and the universal quantifier: one can contextually define $\exists$ and $\forall$ from ϵ.[23] And, at the same time, the Second Epsilon Theorem (see Avigad and Zach, 2024) shows that the epsilon calculus is conservative over first-order logic: an epsilon-term-free first-order formula A is derivable in the epsilon calculus from a set Γ of epsilon-term-free first-order formulas just in case A is derivable from Γ in first-order logic. In that sense, the principles governing ϵ do not seem logically stronger than the logical axioms governing $\forall$ and $\exists$. Note that if it had been assumed that there was always a fact of the matter of what gets denoted by an epsilon term—and hence certain interpretations of ϵ would have been excluded as unintended on non-logical grounds—ϵ would have to be viewed as descriptive rather than logical; but, as explained before, this is not the case here.[24]

[23] E.g., in the first-order case: $\exists x \varphi[x] \leftrightarrow_{df} \varphi[\epsilon x\, \varphi[x]]$ and $\forall x \varphi[x] \leftrightarrow_{df} \varphi[\epsilon x\, \neg\varphi[x]]$. See Avigad and Zach (2024) for further details.

[24] Similarly, and for related reasons, the second-order Axiom of Choice had been regarded as logical, too.

Whether one is willing to come to the same verdict concerning $1d(\mathcal{L}^2_{\in,Set})$ on the basis of Definition 2, too, depends of course on whether one is willing to grant logicality to $Logical\text{-}in\text{-}\mathfrak{C}$. But if there are logical objects, then a concept by which they can be qualified as such should at least count as logical in a slightly extended sense of (*meta-*)logicality. Compare: just as the truth values $true$ and $false$ are logical objects, the concept *truth value* that applies to them should count as a logical concept. In the case of logically true sentences or propositions, the concept of *logical truth* that characterizes them would generally be regarded as a logical concept in a similarly extended sense. And in provability logic, the provability of a logical truth ($Prov(\top)$) is expressed by the same logical operator by which the provability of that provability claim is expressed ($Prov(Prov(\top))$). In the same vein, the concept $Logical\text{-}in\text{-}\mathfrak{C}$ should qualify as a logical concept as well, and hence both Definitions 1 and 2 define $\in$ and Set in properly logical terms.

About (ii): As shown before, the Carnap sentence

(C) $\exists R \exists S\, ZF2[R,S] \rightarrow ZF2[\in, Set]$,

follows logically from Definition 1 in the deductive system of second-order logic extended by the epsilon calculus. Now consider any standard theorem of pure mathematics, such as, say, the Fundamental Theorem of the Calculus (FTC): FTC is known to be logically derivable from $ZF2[\in, Set]$ and suitable explicit set-theoretic definitions (such as of 'real number', 'real function', 'continuous', 'integral', and the like). By the Deduction Theorem, one can therefore logically derive

(D) $ZF2[\in, Set] \rightarrow FTC$

from these definitions. Consequently, if these definitions are combined with Definition 1, one can logically derive from that combination of definitions the sentence

(E) $\exists R \exists S\, ZF2[R,S] \rightarrow FTC$,

as (E) is logically derivable from (C) and (D). This means: even though the standard theorems of pure mathematics, such as FCT, are not quite logically derivable from explicit definitions themselves, one might still say that they are derivable from explicit definitions "in conditional form", that is, as consequents of conditionals in which the Ramsey sentence (R) serves as the antecedent.

In that sense, (quasi-)Carnapian logicism based on Definition 1 (or Definition 2) alone already amounts to a logicist variant of "if-thenism" or deductivism about mathematics,[25] as advocated e.g. in Russell's *Principles of Mathematics* (Russell, 1903). Moreover, even though it is sometimes claimed that Whitehead and Russell's *Principia Mathematica* (Whitehead and Russell, 1910–1910 1913) relied on the Axioms of Choice (or the Multiplicative Axiom) and the Axiom of Infinity—which were of questionable logicist status—what Whitehead and Russell actually suggested was to use these axioms as antecedents of conditionals, such that these conditionals would then be logically derivable in their ramified theory of types (see Whitehead and Russell, 1910–1910 1913, vol. 2, p. 183). The same strategy is employed by Carnap in his *Abriß der Logistik*,

[25] See Paseau and Pregel (2023) for a survey of deductivism.

in which he uses the Axiom of Choice and the Axiom of Infinity as antecedents of logically derivable theorems of simply type theory (see Carnap, 1929, Sections 24b and 24e). In the same manner, one might view the Ramsey sentence (R) as an "Extended Axiom of Infinity"[26] on the condition of which the standard theorems of pure mathematics become logically derivable in the deductive system of second-order logic extended by explicit definitions.

In fact, one can even do a bit better. One might rationally reconstruct mathematical practice as if it were engaged in an all-encompassing conditional proof: let us assume mathematicians suppose the Ramsey sentence (R) with the aim of deriving (given logic and definitions) theorems of pure mathematics, such as FTC from before. Once that has been achieved, one would normally finish such a conditional proof by discharging the assumption and concluding the corresponding conditional, such as (E) above. But now assume that mathematicians never actually get around to discharge their Ramsey sentence assumption but rather continue to work on the (implicit) presupposition that it holds true.[27] If viewed in this way, the proof patterns of actual mathematicians can be rationally reconstructed based solely on the extremely thin logicist grounds of Definition 1 or 2.

I hope this makes transparent why Definitions 1 and 2 are highly attractive for logicist purposes. However, one can still do better: for the Ramsey sentence (R) is not just any old presumption that mathematicians might want to make but may itself be seen to be (likely to be) semantically analytic in a suitable logicist framework. The required notion of semantic analyticity and the relevant logicist framework will be the topic of the next section.

3. Frameworks, Semantic Analyticity, and the Logicist Framework

It is time to shift our attention to the Ramsey sentence of $ZF2[\in, Set]$, that is,

(R) $\exists R \exists S\, ZF2[R, S]$.

Clearly, (R) only consists of logical symbols. If (R) is true, this means it is both a logical sentence and true, which, however, would not mean that (R) is *logically true*. In fact, given our standard Tarskian model-theoretic understanding of logical truth, (R) is of course not logically true, as there are second-order countermodels (e.g. all models with a finite domain). What I want to argue for in the following is that it is nevertheless analytic(ally true) in a suitably defined logicist framework. In contrast with more traditional conceptions of analyticity, such as Kant's, the required Carnapian concept of semantic analyticty-in-a-framework will allow for existence statements to be analytic, about which Carnap is perfectly explicit:[28] e.g., in Carnap (1950), he states that in a suitable arithmetical framework the existence of natural numbers and of prime numbers greater than a million are analytic, of which the former existence claim is trivial while the second one is less so. He also points out that classical logic comes with existence assumptions concerning individual constants such as '5', which may belong to the vocabulary of such a framework; if so, $\exists x\, x = 5$ is analytic because logically true.[29]

[26] Indeed, it is easy to see that if (R) is satisfied by a full second-order model with a first-order domain of a certain cardinality, it is satisfied by every full second-order model with a first-order domain of a greater cardinality. In that sense, if (R) is satisfiable at all, it merely amounts to the claim that there are sufficiently many individuals.

[27] This will match how I am going to describe the attitude of ordinary mathematicians towards (R) in Section 4.

[28] See Ebbs (2017, Chapter 2) for further discussion.

[29] Accordingly, there are possibility formulas in Carnap's (1946) modal predicate logic that are logically true, such as formulas of the form $\Diamond A$, in which A is a contingent non-modal sentence (see Carnap, 1946, p.64). Possibility formulas are the modal counterparts of existence formulas.

But of course it is one thing to acknowledge that Carnap accepted the analyticity of existence statements in certain frameworks and yet another to understand why this might make good philosophical sense. The philosophical point behind this is that conceptual frameworks are meant to organize information by structuring it in a particular manner—information that will then become expressible linguistically by sentences of the thereby interpreted object language of the framework. In that respect, Carnapian frameworks take over some of the structuring roles that space and time had for Kant,[30] though subject to some crucial differences: Kantian intuition of space and time is replaced by the linguistic expression of concepts and propositions; frameworks can be constructed and revised in a great plurality of ways, whilst Kantian space and time are simply given and unrevisable; and a Carnapian conceptual framework does not pre-determine what the empirical world is like, while for Kant e.g. the space of empirical intuition just *is* physical space. As a logical *empiricist*, Carnap regarded it to be the task of empirical science to find out, by observation and experiment, whether a contingent empirical sentence A or its negation $\neg A$ is true of the empirical world. But as a *logical* empiricist, he also thought that A and $\neg A$ are meaningful, and hence they—and with them the rest of their underlying language — come with some abstract conceptual and propositional structure that is constitutive of having thoughts about the world, independently of whether A or $\neg A$ is true of it. This structure needs to be in place prior to empirical investigation, and, once rationally reconstructed, it is that structure that a formal conceptual framework provides and assigns as interpretation to the sentences of its object language. Thereby, a conceptual framework comes itself with an ontological commitment to abstract structured thought. Carnapian logicism suggests to semantically interpret mathematics as dealing precisely with these abstract structured thoughts provided by the conceptual framework itself. The resulting interpretation may even extend to all of standard mathematics as we know it, if only the framework is complex enough, that is, if it provides sufficiently complex relational concepts. In the case of the logicist framework to be introduced below, the Ramsey sentence (R), that is, $\exists R \exists S\, ZF2[R, S]$, is going to express the ontological commitment to such a concept. And the analyticity of (R) in the framework will express that the ontological commitment is provided by the framework itself. Consequently, so long as information about the empirical world is structured according to the rules of the logicist framework to be introduced, each of the trivial classical logical law $A \vee \neg A$, the less trivial definition of membership in Definition 1, and the highly non-trivial Ramsey sentence $\exists R \exists S\, ZF2[R, S]$ will turn out to be (likely to be) true on purely conceptual grounds, independently of what the empirical world is like. In other words: they will be (likely to be) semantically analytic in the framework.[31]

The corresponding Carnapian concept of analyticity-in-a-framework is neither metaphysical nor epistemic in the sense of Boghossian (1996)[32] but rather semantic[33] in exactly the same sense in which Tarski's concept of truth is semantic. In fact, Carnap's definition of analyticity for Language II in his *Logical Syntax* amounts to an early version of a Tarskian definition of

[30] For more on this concerning Kant and time, see Sattig (2025), and for more on the general idea in the context of Carnap's *Aufbau*, see Richardson (1998).

[31] Other than its explicitly semantic formulation, this conception of mathematics is already present in Carnap's *Logical Syntax*. As Friedman (1999, p. 87) formulates it in his "Logical Truth and Analyticity in Carnap's 'Logical Syntax of Language' ": "Mathematics is built in to the very structure of thought and language and is thereby forever distinguished from merely empirical truth."

[32] Metaphysical analyticity is explained in terms of grounding or truthmaking, epistemic analyticity in terms of justification and cognitive grasp of meaning.

[33] I am in agreement with Lavers (2024, p.39) on this point. Otherwise, Lavers' (2024) understanding of, and argument for, the analyticity of large parts of mathematics differ very much from mine. (Lavers' idea is to determine the set of analytic sentences from statements and rules that emerge from the first stage of a Quinean explication.)

truth (see Suppl. G of Leitgeb and Carus, 2024), and once Carnap had fully embraced Tarskian semantics, he presented analyticity by reference to Tarskian semantic rules from the start:

> A sentence S_i is *L-true* in a semantical system S if and only if S_i is true in S in such a way that its truth can be established on the basis of the semantical rules of the system S alone, without any reference to (extra-linguistic) facts (Carnap, 1956, p.10),

where L-truth explicates analyticity, and where a semantical system is nothing but a conceptual framework in our terminology. And just as metaphysical necessity may be described as truth in all metaphysically possible worlds, that is, in all worlds in which the metaphysical laws are held fixed, analyticity-in-a-framework may also be described as truth in all worlds that are semantically possible in the relevant framework, that is, in all worlds in which the semantic rules of the framework are held fixed.

In Carnap's words:

> A sentence S_i is *L-true* (in S_I) $=_{Df}$ S_i holds in every state-description (in S_I) (Carnap, 1956, p.10)

and

> A sentence S is *A-true* in L $=_{df}$ S holds in all admissible models (Carnap, 1963, p. 901)

where A-truth explicates analyticity again.

It is important to note that this notion of analyticity is framework-relative (hence the "*L-true* (in S_I)" and "*A-true* in L"): much as the definition of a mathematical term may differ from one textbook to the next, since different textbooks may organize even the same body of mathematical knowledge differently, a sentence that is analytic in one conceptual framework may well fail to be analytic in another one. That is because the semantic rules of the frameworks may differ, and accordingly the class of semantically possible worlds in one framework may differ from the class of semantically possible worlds of another. Since analyticity in the present Carnapian sense is explicitly defined for, and relative to, constructed artificial frameworks, it is to be distinguished from the notion of analyticity in natural language that was mostly in the forefront of Quine's criticism in "Two Dogmas of Empiricism" (Quine, 1951). But I will not be able to enter the classical Carnap-Quine debate on analyticity here in any more detail.[34] Furthermore, the semantic notion of analyticity in a conceptual framework should be distinguished from metaphysical necessity, too. E.g., if metaphysical necessity got explicated in a conceptual framework with the help of a suitably constructed accessibility relation between worlds, every sentence that is analytically true in the framework would be metaphysically necessary but not necessarily the other way around.[35]

[34] For more on the debate, see Suppl. B of Leitgeb and Carus (2024).

[35] E.g., following Kripkean considerations, the accessibility relation might be constructed in a framework such that there is a semantically possible world at which $Son(a, b) \wedge \neg\Box Son(a, b)$ is true but where there is no metaphysically possible world at which that sentence is true. The reason for constructing a framework like that might be to rationally reconstruct the thought that the sentence does not invalidate any semantic rule but does invalidate the metaphysical necessity of hereditary relationships. ($\Box$ is meant to express metaphysical necessity, and $Son(a, b)$ is meant to express that a is son of b.)

Although Carnap did not use my term 'semantically possible world', he did evaluate formulas relative to entities that represented possible ways the world might be like (possible worlds, possibilities, possible cases, possible states of affairs), such that no semantic rule of the conceptual framework in question would be invalidated by that evaluation.[36]

In particular, Carnap (1942) states general postulates for the notion of the so-called L-range of a formula, by which Carnap's explicates the intensional meaning of a formula, that is, its truth conditions. In §18, he shows how these postulates can be realized by means of different procedures (A, B, C) that define, in a non-extensional metalanguage, L-ranges as classes of propositional entities (so-called L-states). In §19, he does the same for procedures that define L-ranges as classes in an extensional metalanguage: classes of (maximal) state descriptions (procedure E), classes of sentences (procedures F and G), and classes of so-called state-relations (procedures K and L). The state-relations of procedures K and L are similar to models (structures, interpretations) in contemporary model-theory in the sense that they are structured entities of objects and extensional properties/relations of these objects that can then be used to interpret and evaluate sentences.[37] Procedure E is applied later in his *Meaning and Necessity* (Carnap, 1947/1956) in which he presents formulas as holding at state descriptions, such that the (L-)range of a formula is the class of state-descriptions at which the formula holds (Carnap 1947/1956, p. 9). Furthermore, a formula is said to be true simpliciter just in case it holds at the actual state description (Carnap, 1947/1956, p.10); the same idea had been put forward in Carnap (1942, D18-B9) in terms of "rs", that is, "the real L-state". Clearly, this amounts to a precursor of present-day possible worlds semantics in which formulas are evaluated at worlds, one of which is regarded as actual. And in his later work (such as in Carnap, 1963 cited above or in Carnap, 1971), Carnap ends up evaluating formulas relative to models in the contemporary model-theoretic sense.[38]

My notion of semantically possible world in a conceptual framework is but a further development and application of these Carnapian ideas about semantics. So far as the metalanguage is concerned in which I will describe semantically possible worlds and the evaluation of formulas relative to them, I will follow Carnap's semantic work from the 1940s and use a language of higher-order logic instead of first-order set theory. It will be sufficient for my purposes to only sketch that higher-order language and the semantic rules that are formulated within it. The situation will resemble that of a typical logic textbook in which an object language—say, some second-order language—is specified in full formal detail, whereas the metalanguage in which the semantic rules for that object language are formulated remains partially unspecified (although a full formal specification could be given in principle).

Now let me turn to the logicist conceptual framework $\mathfrak{C}$, which involves the following components:

[36] See Suppl. F of Leitgeb and Carus (2024) for more on Carnap's intensional semantics.

[37] There are also differences: unlike models of modern model theory, which assigns e.g. a class of objects to each unary *predicate* of the object language, a state-relation in Carnap's procedure L assigns a class of objects to each extensional *property* that is to be expressed in the object language. Moreover, where modern model theory describes models in the language of standard first-order set theory, Carnap (1942) describe state-relations in the language of higher-order logic (type theory).

[38] The fact that *Meaning and Necessity* (Carnap, 1947/1956) presented (L-)ranges as classes of state descriptions, and thus of syntactic entities, is sometimes interpreted as if Carnap had not left behind the syntactic emphasis of his *Logical Syntax* and hence had not fully embraced possible worlds semantics as yet. But that would be a misinterpretation: as he explains in Footnote 9 on p. 9 of Carnap (1947/1956), he only opted for applying procedure E from *Introduction to Semantics* because it seemed "the most convenient" one for the purposes of *Meaning and Necessity*. But other than that he might just as well have opted for a non-syntactic reconstruction of possible worlds, as witnessed by procedures K and L in Carnap (1942). I am grateful to Pierre Wagner for a discussion of these points.

(i) a second-order object language $\mathcal{L}$ with the usual primitive logical symbols of second-order logic, the additional primitive logical symbols ϵ and *Logical-in-*$\mathfrak{C}$, the defined predicates $\in$ and Set, and (for merely illustrative purposes) the primitive descriptive unary predicates 'Man' and '$Married$', and the defined unary predicate '$Bachelor$' (the last three predicates are tacitly relativized to a fixed point of time);

(ii) semantic rules for $\mathcal{L}$, formulated in a metalanguage of (cumulative) higher-order logic with ϵ, the logical predicate *Logical-in-*$\mathfrak{C}$, syntactic terms concerning the syntax of $\mathcal{L}$, the primitive descriptive unary predicates 'Man' and '$Married$', and some optional further expressions to be described in Section 4; the axioms and rules of a suitable deductive system of higher-order logic with extensionality and the epsilon calculus governing that metalanguage; and some further postulates, such as the definitions to be presented below and some optional additional postulates to be described in Section 4;

(iii) a class $\mathfrak{W}$ of models $\mathfrak{M}$, such that (iii.i) every way of assigning extensions to the primitive descriptive terms of $\mathcal{L}$ is realized by a (uniquely determined) $\mathfrak{M}$ in $\mathfrak{W}$, and (iii.ii) truth in each of these models $\mathfrak{M}$ respects the semantic rules for $\mathcal{L}$ in (ii).

(ii) means that $\mathfrak{C}$ involves some metalinguistic deductive components, whilst (iii) means that it also includes semantic components.

$\mathfrak{W}$ is of course the class of all semantically possible worlds of the framework $\mathfrak{C}$, which results from running through all combinatorial possibilities of assigning extensions to the primitive descriptive expressions in $\mathcal{L}$. Since all combinatorial possibilities are realized in $\mathfrak{W}$, it will be guaranteed that one of the worlds in $\mathfrak{W}$ corresponds to the actual world: it is the world at which, e.g., the extension of $Married$ is indeed the class of married humans at the fixed point of time; etc. (See the definition below.)

In contrast, a (proper) theory in $\mathfrak{C}$ would be given semantically by a proper subclass of $\mathfrak{W}$. Thus, unlike $\mathfrak{W}$, theories in $\mathfrak{C}$ rule out at least one semantically possible world in $\mathfrak{C}$, which is also why they are not guaranteed to include the actual world.

Finally, note that the semantically possible worlds of $\mathfrak{C}$ do not have any world-relative first-order or second-order domains assigned to them.

I will not state all of the semantic rules of $\mathfrak{C}$ for $\mathcal{L}$, but they include:

For all $\mathfrak{M}$ in $\mathfrak{W}$, for all variable assignments s:[39]

$Val_{\mathfrak{M},s}(Married(x)) = 1$ iff $\mathfrak{M}(Married)(s(x))$.
$Val_{\mathfrak{M},s}(Man(x)) = 1$ iff $\mathfrak{M}(Man)(s(x))$.
$Val_{\mathfrak{M},s}(Bachelor(x)) = 1$ iff not $\mathfrak{M}(Married)(s(x))$ and $\mathfrak{M}(Man)(s(x))$.
$Val_{\mathfrak{M},s}(Logical\text{-}in\text{-}\mathfrak{C}(x)) = 1$ iff $Logical\text{-}in\text{-}\mathfrak{C}(s(x))$.
$Val_{\mathfrak{M},s}(x \in y) = 1$ iff $Val_{\mathfrak{M},s}(\epsilon R \exists S\, ZF2[R,S])(s(x), s(y))$.
[$Val_{\mathfrak{M},s}(x \in y) = 1$ iff $Val_{\mathfrak{M},s}(\epsilon R \exists S(\forall z(S(z) \rightarrow Logical\text{-}in\text{-}\mathfrak{C}(x)) \wedge ZF2[R,S])(s(x), s(y))$.]
$Val_{\mathfrak{M},s}(Set(x)) = 1$ iff $Val_{\mathfrak{M},s}(\exists y\, x \in y) = 1$.
$Val_{\mathfrak{M},s}(S(x)) = 1$ iff $s(S)(s(x))$.
$Val_{\mathfrak{M},s}(R(x,y)) = 1$ iff $s(R)(s(x), s(y))$.
$Val_{\mathfrak{M},s}(\neg\varphi) = 1$ iff not $Val_{\mathfrak{M},s}(\varphi) = 1$.

[39] In the present context, any talk of quantification over variable assignments s is short for: talk of second-order quantification over functions that map first-order variables to individuals, and talk of third-order quantification over functions that map second-order variables to second-order entities.

$Val_{\mathfrak{M},s}(\varphi \wedge \psi) = 1$ iff $Val_{\mathfrak{M},s}(\varphi) = 1$ and $Val_{\mathfrak{M},s}(\psi) = 1$.
$Val_{\mathfrak{M},s}(\forall x\, \varphi) = 1$ iff for all x-alternatives s' of s it holds: $Val_{\mathfrak{M},s'}(\varphi) = 1$.
$Val_{\mathfrak{M},s}(\forall R\, \varphi) = 1$ iff for all R-alternatives s' of s it holds: $Val_{\mathfrak{M},s'}(\varphi) = 1$.
$Val_{\mathfrak{M},s}(\epsilon R\, \varphi) = \epsilon s'(s'$ is an R-alternative of s and $Val_{\mathfrak{M},s'}(\varphi) = 1)(R)$.

As usual, the semantic rules determine uniquely, for each $\mathfrak{M}$ and s, an evaluation function $Val_{\mathfrak{M},s}$ that maps formulas in $\mathcal{L}$ to truth values. The evaluation of the formula $Married(x)$ at $\mathfrak{M}$ depends on what worldly extension $\mathfrak{M}(Married)$ the world $\mathfrak{M}$ assigns to the primitive descriptive predicate $Married$; analogously for $Man(x)$. The semantic rule for $Bachelor(x)$ encodes the definition of the defined descriptive predicate $Bachelor$ as applying precisely to unmarried men; since $Bachelor$ is defined from $Married$ and Man in $\mathfrak{C}$, its world-relative extension varies with those of $Married$ and Man. The semantic rules for the object-linguistic formula $Logical\text{-}in\text{-}\mathfrak{C}(x)$ invokes the meta-linguistic formula $Logical\text{-}in\text{-}\mathfrak{C}(x)$ (much as the semantic rule for $\neg$ involves 'not'). The semantic rules for $\in$ and Set encode Definition 1 (or Definition 2) from Section 2. The semantic rules for atomic formulas with a class variable S or a relation variable R are standard, as are those for the usual logical symbols. Finally, the semantic rule for object-linguistic epsilon terms $\epsilon R\, \varphi$ employs a metalinguistic epsilon term of the form $\epsilon s'(\ldots)$ in which s' is a variable for functions.

On that basis, we can define various further semantic notions well-known from intensional semantics: e.g., the proposition expressed by a sentence φ of $\mathcal{L}$ in $\mathfrak{C}$ is the class of semantically possible worlds $\mathfrak{M}$ in $\mathfrak{W}$, such that for all s, $Val_{\mathfrak{M},s}(\varphi) = 1$. The concept expressed by the unary predicate $Married$ of $\mathcal{L}$ in $\mathfrak{C}$ is the function that maps each world $\mathfrak{M}$ in $\mathfrak{W}$ to the extension $\mathfrak{M}(Married)$ at $\mathfrak{M}$. Etc.

Moreover, we can define actuality(-in-$\mathfrak{C}$), the metalinguistic semantic predicate 'true(-in-$\mathfrak{C}$)', and the metalinguistic semantic predicate 'analytic(-in-$\mathfrak{C}$)'. A world is actual just in case it assigns the "right" or intended extensions to all primitive descriptive predicates of the object language $\mathcal{L}$, as can be captured by translating these predicates into the metalanguage. Truth (simpliciter) of a sentence in $\mathcal{L}$ is its truth at the actual world (one can prove there is only one), while a sentence in $\mathcal{L}$ is analytic just in case it holds at all semantically possible worlds in $\mathfrak{C}$:

Metadefinition 3. (Actuality-in-$\mathfrak{C}$)

For all $\mathfrak{M}$ in $\mathfrak{W}$:

$\mathfrak{M}$ is actual(-in-$\mathfrak{C}$) iff for all d:
$d \in \mathfrak{M}(Married)$ iff d is married (at the given fixed point in time), and
$d \in \mathfrak{M}(Man)$ iff d is a man (at the given fixed point in time).

Metadefinition 4. (Truth-in-$\mathfrak{C}$)

For all sentences φ in the object language $\mathcal{L}$ of $\mathfrak{C}$:

φ is true(-in-$\mathfrak{C}$) iff
for all $\mathfrak{M}$ in $\mathfrak{W}$, for all s: if $\mathfrak{M}$ is actual(-in-$\mathfrak{C}$), then $Val_{\mathfrak{M},s}(\varphi) = 1$.

Metadefinition 5. (Analyticity-in-$\mathfrak{C}$)

For all sentences φ in the object language $\mathcal{L}$ of $\mathfrak{C}$:

φ is analytic(-in-$\mathfrak{C}$) iff for all $\mathfrak{M}$ in $\mathfrak{W}$, for all s: $Val_{\mathfrak{M},s}(\varphi) = 1$.

Hence, if a sentence of $\mathcal{L}$ is analytic(-in-$\mathfrak{C}$), it neither rules out any assignment of extensions to primitive descriptive terms in $\mathcal{L}$ nor any theory in $\mathfrak{C}$. For the same reason, an analytic sentence in the framework is guaranteed to be true at the actual world of the framework.

Here are some analytic example sentences in $\mathcal{L}$, the analyticity of which can be logically derived from the semantic rules of $\mathfrak{C}$:

$\forall x(Married(x) \vee \neg Married(x))$ is analytic(-in-$\mathfrak{C}$).
$\forall x(Bachelor(x) \leftrightarrow \neg Married(x) \wedge Man(x))$ is analytic(-in-$\mathfrak{C}$).
$\forall x, y(x \in y \leftrightarrow (\epsilon R \exists S\, ZF2[R, S])(x, y))$ is analytic(-in-$\mathfrak{C}$).
$\forall x(Set(x) \leftrightarrow \exists y\, x \in y)$ is analytic(-in-$\mathfrak{C}$).
$\exists R \exists S\, ZF2[R, S] \rightarrow \exists S\, ZF2[\epsilon R \exists S\, ZF2[R, S], S]$ is analytic(-in-$\mathfrak{C}$).

Thus, e.g., both parts of Definition 1 from Section 2 reappear as object-linguistic statements in $\mathcal{L}$ that are analytic(-in-$\mathfrak{C}$). The same applies to Definition 2 if the corresponding alternative semantic rule for $x \in y$ from above is used.

More generally, all logical axioms of the deductive system of second-order logic formulated in the object language are semantically analytic, and the same holds for all explicit definitions formulated in the object language and for all axioms of the epsilon calculus in the object language. Since Metadefinition 3 also clearly implies that semantic analyticity is closed under logical derivability, all Frege-analytic sentences in $\mathfrak{C}$ are therefore semantically analytic in $\mathfrak{C}$, as promised.

However, this does not mean that *every* sentence in $\mathcal{L}$ is such that it is analytic or its negation is analytic in $\mathfrak{C}$. E.g.:

$\exists x\, Bachelor(x)$ is not analytic(-in-$\mathfrak{C}$).
$\neg \exists x\, Bachelor(x)$ is not analytic(-in-$\mathfrak{C}$).

The reason for this is that there are semantically possible worlds in $\mathfrak{C}$ at which the extension of Man is a subclass of the extension of $Married$ and hence there are no bachelors, and there are semantically possible worlds in $\mathfrak{C}$ at which this is not the case and so there are bachelors. Similarly, a sentence expressing that there are exactly 1000 bachelors would not be analytic in $\mathfrak{C}$, and its negation would not be analytic in the framework either. This is just as intended: the truth or falsity of these claims does not just depend on the framework but also on the empirical facts; that is: it does not just depend on how information is structured in the framework but also on what information the actual world provides. Accordingly, some theories in the framework are going to claim that there are exactly 1000 bachelors, others that there are not, and yet others are going to claim neither. It is a matter of empirical investigation to confirm or disconfirm such theories, but all of these theories would be formulated against the backdrop of the framework $\mathfrak{C}$. If the relevant point of time is right now, we know in fact on empirical grounds that there are not exactly 1000 bachelors, so any object language sentence saying so is true at the actual world. Moreover, if $Bachelor$ had not been defined as applying to all and only unmarried men but had been regarded as primitive in $\mathfrak{C}$, the extension of $Bachelor$ would have varied independently of those of $Married$ and Man in the corresponding alternative framework $\mathfrak{C}'$. Hence, $\forall x(Bachelor(x) \leftrightarrow \neg Married(x) \wedge Man(x))$ would *not* have been analytic(-in-$\mathfrak{C}'$), since information would have been organized differently in $\mathfrak{C}'$ than in $\mathfrak{C}$.

Now let us return to our Ramsey sentence (R), which is a member of both the object language $\mathcal{L}$ of $\mathfrak{C}$ and of the metalanguage of $\mathcal{L}$ (that metalanguage also belongs to $\mathfrak{C}$).

The semantic rules of $\mathfrak{C}$ yield for all $\mathfrak{M}$ in $\mathfrak{W}$ and for all s:

$Val_{\mathfrak{M},s}(\exists R \exists S\, ZF2[R,S]) = 1$ iff
there is an R/S-alternative s' of s, such that $Val_{\mathfrak{M},s'}(ZF2[R,S]) = 1$ iff
there are an R and an S, such that $ZF2[R,S]$.

Using this, it follows:

$\exists R \exists S\, ZF2[R,S]$ is analytic(-in-$\mathfrak{C}$) iff
for all $\mathfrak{M}$ in $\mathfrak{W}$, for all s: $Val_{\mathfrak{M},s}(\exists R \exists S\, ZF2[R,S]) = 1$ iff
there are an R and an S, such that $ZF2[R,S]$.

The analytic truth of (R) in $\mathfrak{C}$ therefore boils down to a satisfiability claim,[40] that is, to the existence of higher-order R and S satisfying $ZF2[R,S]$. This result is a consequence of the definition of analyticity(-in-$\mathfrak{C}$), the fact that (R) only includes logical symbols, and the semantic rules of $\mathfrak{C}$. In particular, the semantic interpretation of the logical symbols is the same at all worlds, and the semantic rules in $\mathfrak{C}$ for existence claims do not invoke world-relative domains that would restrict the range of existential quantifiers. That is why the reference to worlds $\mathfrak{M}$ has dropped out from the evaluation of (R) once the semantic clauses have been fully unpacked. The analyticity of (R) in $\mathfrak{C}$ therefore follows to consist in the metalinguistic translation of (R) being the case.

More generally, if a sentence φ in $\mathcal{L}$ only includes logical symbols, then for all $\mathfrak{M}$ in $\mathfrak{W}$ and for all s it holds:

if $Val_{\mathfrak{M},s}(\varphi) = 1$ then φ is analytic(-in-$\mathfrak{C}$), and
if $Val_{\mathfrak{M},s}(\varphi) = 0$ then $\neg\varphi$ is analytic(-in-$\mathfrak{C}$).

Consequently, every logical sentence φ is analytic(-in-$\mathfrak{C}$) or its negation $\neg\varphi$ is analytic(-in-$\mathfrak{C}$), which is just as what Carnap had proved for all closed logical formulas of his languages I and II of his *Logical Syntax* (see Carnap 1934/1937, Theorems 14.3 and 34e.11).[41] This does not mean, of course, that for all logical φ, either φ is derivable from the deductive components of the framework $\mathfrak{C}$ or its negation $\neg\varphi$ is; after all, analyticity has been defined semantically, not proof-theoretically. It only means that purely logical statements are such that, if true, they are analytically true, and if false, they are analytically false.

The semantic rules for quantification in $\mathfrak{C}$ may be viewed as either tacitly assigning for each type one and the same domain to the quantifiers in $\mathcal{L}$ at all worlds, or as interpreting the quantifiers in $\mathcal{L}$ unrestrictedly, that is, as quantifying over everything of the right type—everything there is of that type (as expressed by the corresponding metalinguistic universal and existential quantifier).[42]

Indeed, for much of his work, Carnap himself used a "one-domain assumption" (cf. Hintikka, 1991, but see also Schiemer, 2013), and quantification over "absolutely everything" has been shown to be coherent if the semantic rules are formulated using the resources of higher-order logic (Williamson, 2003, see). Moreover, Linsky and Zalta (1994) and Williamson (1998) have advocated the analogous usage of possible worlds semantics with a single universal first-order domain for the interpretation of metaphysical modalities.

[40] This bears some similarity to Hilbert's views on mathematical truth and consistency: "if the arbitrarily given axioms do not contradict one another. . . then they are true and the things defined by the axioms exist" (Hilbert, 1899, p.39).
[41] Thus, if R is false in $\mathfrak{C}$, it is analytically false in $\mathfrak{C}$, i.e., its negation is analytic in $\mathfrak{C}$.
[42] But note that what there is does not necessarily exhaust what is metaphysically possible.

Either way, since the worlds in $\mathfrak{W}$ are meant to track variations in extensional interpretation and not variations of what exists, it should be fair enough not to vary the ranges of quantifiers with worlds. Even more importantly for present purposes, Carnapian tolerance should allow us to set up our logicist framework as we please, so long as it may still count as logicist. And the world-independent interpretation of quantifiers in $\mathfrak{C}$ certainly does not undermine any logicist tenets.

So where does this leave us with the analyticity of the object-linguistic Ramsey sentence (R) in $\mathcal{L}$? It leaves us with the follow-up question

(Q) Are there R and S, such that $ZF2[R, S]$?

which is formulated in the metalanguage of $\mathcal{L}$ that also belongs to our logicist framework $\mathfrak{C}$. As shown before, if the answer to (Q) is 'yes', (R) will be analytic (-in-$\mathfrak{C}$), hence $ZF2[\in, Set]$ will be analytic(-in-$\mathfrak{C}$), and thus also part 2d($ZF2[\in, Set]$) of our logicist thesis from Section 1 will be vindicated.

In the next section I am going to argue that the answer to (Q) is indeed likely to be 'yes', which is why $ZF2[\in, Set]$ is likely to be analytic in $\mathfrak{C}$.

4. The (Likely) Analyticity of the Ramsey Sentence

One way of settling question (Q) from the last section would be by brute force: one might simply assume the metalinguistic translation of the Ramsey sentence (R) to be included in the metalinguistic deductive components of our logicist framework $\mathfrak{C}$, by which the analyticity of the object-linguistic Ramsey sentence (R) in $\mathfrak{C}$ would become derivable in $\mathfrak{C}$.

While this might seem a bit like cheating, there would be nothing in principle wrong about doing so. This said, there are three reasons for which I am nevertheless not going to pursue that strategy: first, we are only searching for an answer to (Q), not a provable answer. Put another way: the mere existence of an R and S satisfying $ZF2[R, S]$ is sufficient for (R) being analytic(-in-$\mathfrak{C}$). Therefore, while *proving* that existence claim would conveniently deliver the existence of such R and S, it would also go beyond what is required.[43] Second, consider anyone who might still question (perhaps on Quinean holistic grounds) the viability of distinguishing between the conceptual framework $\mathfrak{C}$ and the proper theories in $\mathfrak{C}$, as presented in the last section: any such person would surely feel only more concerned if $\mathfrak{C}$ were to include deductively strong components, such as the metalinguistic translation of (R). And third, the stronger the deductive components of a conceptual framework, the greater the risk of the framework being inconsistent, and inconsistency would be just as unattractive to the constructor of a Carnapian framework based on classical logic as it would be to anyone putting forward a scientific theory based on classical logic. So I refrain from building (R) into the framework deductively: I will leave the deductive components of the framework $\mathfrak{C}$ as deductively weak as they were described in the last section, consisting just of semantic rules, a deductive system of logic, and explicit definitions.[44]

Instead, I suggest conducting the following little thought experiment: *what if one presented the conceptual framework $\mathfrak{C}$ from the last section to ordinary mathematicians and set theorists?* One would

[43] Compare the related discussion in Awodey and Carus (2003, 2004), who point out against Gödel that a Carnapian framework based on classical logic does not have to be *provably* consistent, just consistent.

[44] I am grateful to an anoymous reviewer for urging me to comment on this point.

explain to them that the quantifiers in (R) are meant to range over everything of the right type, or that there is a fixed intended universe of discourse that is tacitly meant to include all of the usual mathematical entities of the right type. And then one would pose to them question (Q) as a logical-mathematical question:

(Q) Are there R and S, such that $ZF2[R, S]$?

In their roles as experts for such logical-mathematical questions, *what would they answer?*

I take it that most ordinary mathematicians accept or presuppose $ZF2[\in, Set]$ as a coherent interpreted background language that has never led to contradictions and which they find more or less conducive to their own mathematical work—work that does not itself concern models of set theory but rather number-theoretic properties of integers, probabilistic properties of random walks in graphs, fixed-point properties of continuous functions on topological spaces, and the like. For that reason, they should be willing to accept or presuppose (R), too, as (R) is logically entailed by $ZF2[\in, Set]$ in the deductive system of second-order logic, and they have been willing to accept or presuppose $ZF2[\in, Set]$ as a foundation. If they were forced to comment more particularly on the existence of *set*-sized models of $ZF2[\in, Set]$ and hence to comment on the existence of *set* values of R and S in 'there are R and S, such that $ZF2[R, S]$' (rather than proper-class-sized entities), they might point out: no one knows conclusively whether such a set model exists, as it seems that we can neither derive (R) nor its negation from uncontroversial principles. After which they might defer to the experts on set models, that is, their set theorist colleagues.

In turn, set theorists do study models of set theory. And they do have more to say about the existence of models of $ZF2[\in, Set]$: they might put forward the established result that if there is a strongly inaccessible cardinal greater than ω, then there is a set model of $ZF2[\in, Set]$. And at least those set theorists (called "absolutist practitioners" in Kant, 2025,?) who believe in the existence of a uniquely determined universe of sets that makes certain set-theoretic axioms true would voice their belief in the existence of such strongly inaccessible cardinals.[45] And they might give arguments for this, too, even when these arguments could not be formally reconstructed as proofs in $ZF2[\in, Set]$ or first-order ZFC (assuming these theories to be consistent, as set theorists very strongly believe them to be).[46] So at least "absolutist" set theorists would not just answer (Q) with a 'yes', they would even think the witnesses to 'there are R and S, such that $ZF2[R, S]$' may be taken to be sets. Of course, they might still be wrong about all of that—after all, no deductively valid argument with obviously true premises has been put forward. But there still seem to be an *inductively strong arguments* (in the sense of Skyrms, 2000, p.17) in favor of (R): arguments that make (R) likely or plausible. Just as all other inductively strong arguments, they do not guarantee the truth of their conclusion given

[45] See Kant (2025, 81–3) and Kant (2025, 114), who examined this empirically, and who reports that absolutist practitioners believe in the truth of large cardinal axioms, at least up to Woodin cardinals (and thus including strongly inaccessible cardinals). Moreover, set theorists in general widely use large cardinal axioms (Kant, 2025, p. 110), and they believe large cardinal axioms are consistent (Kant, 2025, p.113). (I am very grateful to Deborah Kant for her help on this matter.) Džamonja (2017, Section 3) comments on large cardinals in a similar manner: "Not only are the large cardinals needed for set theory but they are also known to be needed for some seemingly innocent statements about number theory. For example, Harvey Friedman [. . .] developed the Boolean relation theory, which demonstrates the necessity of large cardinals for deriving certain propositions considered "concrete". Friedman and others view this as an obvious reason for a working mathematician to accept large cardinals."

[46] See Hrbacek and Jech (1999, pp.279f) for such an argument. Kant (2025) also makes the point that even set theorists who are finally interested in first-order ZFC proofs (such as in descriptive set theory) regularly use large cardinal axioms and then eliminate them in their proofs. This may be viewed as an argument for the thesis that the assumption of the existence of a strongly inaccessible cardinal is at least instrumentally acceptable for these set theorists.

their premises, but that does not mean that they do not supply any justification whatsoever, and arguing inductively may well be the best we can do at that foundational level.

Summing up: I think it is fair to say that what the verdicts of the experts—ordinary mathematicians and set-theorists—would reveal about their beliefs about (R) in our little thought experiment can be rationally reconstructed as a *high-probability assignment to* (R). Given that, it must be at least as likely that the Ramsey sentence (R) is analytic(-in-$\mathfrak{C}$). I am going to make this probabilistic reconstruction a bit more precise now. Afterwards, I will address two potential worries about the thought experiment.

So far as ordinary mathematicians are concerned, their mathematical statements may best be reconstructed as made *from within our framework* $\mathfrak{C}$ and hence as belonging to the object language $\mathcal{L}$ of $\mathfrak{C}$. The mathematicians' belief or acceptance of such statements may then be reconstructed by means of subjective probability measures that assign probabilities to the members of $\mathcal{L}$. Accordingly, in Carnap's work on inductive logic (see e.g. Carnap, 1950), a conceptual framework such as our $\mathfrak{C}$ is expanded by a corresponding class of such subjective probability measures—say, the class $Prob_{\mathfrak{C}}$—precisely for the purpose of capturing rational inductive reasoning that takes place internally to the framework. And what was said above about mathematicians generally accepting or presupposing $ZF2[\in, Set]$ and hence (R) will then correspond to: for all P in $Prob_{\mathfrak{C}}$ it holds that $P(R) = 1$. That is: for mathematicians speaking from within the framework it is not an epistemic possibility that (R) fails, since for them (R) is epistemically presupposed in their mathematical work and hence must be counted as (group-subjectively) probabilistically certain. If (R) is indeed analytic-in-$\mathfrak{C}$, this intended probabilistic reconstruction will automatically follow from (R) being true in every semantically possible world in $\mathfrak{C}$, and from the probability of a sentence A of the object language of $\mathfrak{C}$ corresponding to the probability of the class of semantically possible worlds of $\mathfrak{C}$ in which A is true (see Carnap, 1971).[47] However, for the same reason, we cannot extract much of an argument in favor of (R) from the ordinary mathematicians' verdicts about (R) other than they are willing to accept or presuppose (R) in their mathematical work.

Now for the rational reconstruction of what is conveyed by the set-theorists' verdicts: their statements may be reconstructed as belonging to the metalanguage of the object language $\mathcal{L}$ of our framework $\mathfrak{C}$, as they are reflecting on models of mathematics and set theory. The beliefs or acceptances that these statements express should thus be captured by subjective probability measures that assign probabilities not to the sentences of the object language $\mathcal{L}$ but of the metalanguage of $\mathcal{L}$ (the same language in which analyticity-for-$\mathcal{L}$ had been defined). Since we have seen set theorists would generally answer (Q) with a reasonably strong affirmation based on inductively strong plausibility arguments, their answer may be rationally reconstructed as expressing a high (group-subjective) probability claim of the form *it is likely that there are* R *and* S*, such that* $ZF2[R, S]$. And since it seems rational to defer to the experts on that subject matter, our own rational degrees of belief should concur.

On that basis, summarized in slightly compressed terms, we get the following informal and partially *probabilistic* metalinguistic argument for (quasi-)Carnapian logicism, in which '$P(A) = \ldots$' is a rational-degree-of-belief operator applicable to the sentences A of the metalanguage of $\mathcal{L}$

[47] Indeed, a sentence A in the object language $\mathcal{L}$ of framework $\mathfrak{C}$ might be defined to be *apriori relative to* $\mathfrak{C}$ just in case for all probability measures P in $Prob_{\mathfrak{C}}$ it holds that $P(A) = 1$. I regard this as a suitable rational reconstruction of the epistemic notion of relative or constitutive aprioricity discussed by Friedman (2001) (amongst others), but I will not defend this claim here. Note that every sentence that is analytic-in-$\mathfrak{C}$ is also apriori relative to $\mathfrak{C}$ but not necessarily the other way around.

in our logicist conceptual framework $\mathfrak{C}$, such that the respective subject whose rational degrees of belief are denoted by '$P(\cdot)$' is *us*.[48] '$Analytic$' is short for 'analytic(-in-$\mathfrak{C}$)', 'ε' denotes some small but only vaguely determined number, I will suppress all matters to do with quotation, and I will concentrate just on Definition 1 and (R) again:

(a) $(\in =_{df} \epsilon R \exists S\, ZF2[R,S]) \wedge \forall x(Set(x) \leftrightarrow_{df} \exists y\, x \in y)$.
(b) $P(Analytic((\in = \epsilon R \exists S\, ZF2[R,S]) \wedge \forall x(Set(x) \leftrightarrow_{df} \exists y\, x \in y))) = 1$.
(c) $P(Analytic(\exists R \exists S\, ZF2[R,S] \rightarrow \exists S\, ZF2[\epsilon R \exists S\, ZF2[R,S], S])) = 1$.
(d) $P(\exists R \exists S\, ZF2[R,S] \leftrightarrow Analytic(\exists R \exists S ZF2[R,S])) = 1$.
(e) $P(\exists R \exists S\, ZF2[R,S]) = 1 - \varepsilon$.
(f) $P(Analytic(\exists R \exists S\, ZF2[R,S])) = 1 - \varepsilon$.
(g) $P(Analytic(\exists S\, ZF2[\epsilon R \exists S\, ZF2[R,S], S])) \geq 1 - \varepsilon$.
(h) $P(Analytic(ZF2[\in, Set])) \geq 1 - \varepsilon$.
(i) $(\in =_{df} \epsilon R \exists S\, ZF2[R,S]) \wedge \forall x(Set(x) \leftrightarrow_{df} \exists y\, x \in y) \wedge\ P(Analytic(ZF2[\in, Set])) \geq 1 - \varepsilon$

Therefore, (quasi-)Carnapian logicism holds.

(a) is Definition 1 from Section 2, but now viewed as a metalinguistic statement that says correctly how the object-linguistic terms $\in$ and Set in $\mathcal{L}$ have been defined in $\mathfrak{C}$. Since Definition 1 is provably analytic in $\mathfrak{C}$, as shown in Section 3, (b) rightly states that the subjective probability that Definition 1 is analytic(-in-$\mathfrak{C}$) is 1 (since we are certain that (a) is the case). The same holds for the subjective probability of the analyticity of the Carnap sentence (C) in (c). (d) reflects it being provable in $\mathfrak{C}$ that the analyticity of (R) boils down to the metalinguistic translation of (R), as demonstrated in Section 3. (e) is the rational reconstruction of our deference to our set theorists' informed verdicts about that metalinguistic translation of (R). (f) follows from (d) and (e) by the axioms of probability. (g) follows from (c) and (f) together with the logical closure of analyticity (shown in Section 3 and us being certain of it) and the axioms of probability. Similarly, (h) follows from (g), (b), the logical closure of analyticity, and the axioms of probability. (i) just joins (a) and (h) by conjunction.

But (i) yields the promised thesis of (quasi-)Carnapian logicism from Section 1, since $\mathfrak{C}$ is a framework in which all mathematical terms in $\mathcal{L}^2_{\in,Set}$ are logical in $\mathfrak{C}$ (1d($\mathcal{L}^2_{\in,Set}$), and all mathematical theorems of $ZF2[\in, Set]$ are likely to be analytic in $\mathfrak{C}$ (2d($ZF2[\in, Set]$.

Let me conclude by addressing two potential worries about our previous little thought experiment of asking mathematicians and set theorists about (R)—one epistemological, the other one ontological. The epistemological one is: is it permissible for a logicist about mathematics to justify a statement by asking mathematicians for their opinion about it? Wouldn't that be viciously circular? And the ontological worry is: the reason set theorists believe there to be R and S that satisfy $ZF2[R,S]$ is that they strongly believe there to be a relation of *sets* and a class of *sets* that jointly satisfy $ZF2[R,S]$. But what reason do we have to believe these sets are logical objects, as one might perhaps require of a logicism about mathematics?

A brief inspection of the logicist thesis I promised to defend in Section 1 should swiftly clarify that it is neither epistemological nor ontological in nature but rather semantic: 1c was about terms and their meaning in a framework, whilst 2c was about theorems and their semantic

[48] Carnap (1950, 1971) would have rationally reconstructed such a probabilistic argument in yet another conceptual framework $\mathfrak{C}^*$, so that the metalanguage in $\mathfrak{C}$ would have become the object language $\mathcal{L}^*$ in $\mathfrak{C}^*$. And then he would have considered logical probability measures that would assign probabilities to the members of $\mathcal{L}^*$. Subjective probability measures such as P would have resulted from conditionalizing such logical probability measures on the available evidence. But I will not be able to go into any more detail on this. See Sznajder (2018) for more on Carnap on inductive probability.

analyticity in a framework. And the existence of a logicist framework in which 1c and 2c are the case was promised to follow from the existence of a logicist framework in which the claims 1d($\mathcal{L}^2_{\in,Set}$) and 2d($ZF2[\in, Set]$) are the case, which are equally semantic. Thus, neither the epistemological nor the ontological worry expressed before actually concerns the logicist project of this paper.

In particular, while e.g. Frege's logicism was certainly at least partially motivated by epistemological concerns, (quasi-)Carnapian logicism is not. Of course, the successful rational reconstruction of a scientific theory may occasionally improve the epistemic standing of that theory. But a logicism about mathematics that proceeds by logically reconstructing the axiomatic theory $ZF2[\in, Set]$, which may itself be viewed as having resulted from the set-theoretic rational reconstruction of mathematical practice, would be extremely unlikely to stand on better justified grounds than $ZF2[\in, Set]$ itself. And indeed none of this is the point of (quasi-)Carnapian logicism, and it has not been claimed to be so either. On the contrary, a (quasi-) Carnapian logicist may happily admit that logic and set theory are epistemologically on par, which is why asking our set theory experts for their advice on a higher-order existence statement should hardly count as a no-go.[49]

With respect to the ontological worry from above, (quasi-)Carnapian logicism is not affected by it because its logicist thesis only concerns the logicality of mathematical terms and the analyticity of mathematical theorems, not the logicality of mathematical objects. As mentioned in Section 2, its application to quasi-categorical second-order set theory only cares about logical structure, not what the entities are like that are structured as such. Accordingly, it does not matter whether the witnesses to 'there are R and S, such that $ZF2[R, S]$' are physical entities, mental entities, proper relations/classes of sets, or quite simply sets, so long as the object-linguistic Ramsey sentence (R) comes out as analytic(-in-$\mathfrak{C}$).

This said, one might also consider a variant of (quasi-)Carnapian logicism that would expand its focus beyond mathematical terms and theorems to *objects*: the corresponding extended logicist theses would still start with

> There is a logicist conceptual framework, such that [. . .]

but then extend 1c and 2c by

3c. all standard mathematical objects are logical objects in the framework,

and extend 1d($\mathcal{L}^2_{\in,Set}$) and 2d($ZF2[\in, Set]$) by

3d(Set). all members of Set are logical objects in the framework.

That is where Definition 2 from Section 2 comes in handy: assume the explicit epsilon term definition of $\in$ (and indirectly of Set) in $\mathfrak{C}$ to include the restriction to objects that are $Logical\text{-}in\text{-}\mathfrak{C}$. And consider the members of the class $Logical\text{-}in\text{-}\mathfrak{C}$ to be abstract objects introduced by the framework $\mathfrak{C}$ itself. This would be in line with how Carnap's (1950) "Empiricism, Semantics, and Ontology" describes what it takes for a framework to introduce a new class of abstract objects: the framework needs to provide a general term for these objects

[49] This deference to set theorists only pertains to the existence of R and S satisfying $ZF2[R, S]$, not to any philosophical thesis of logicism about mathematics. Set theorists are experts concerning the former but not concerning the latter.

($Logical\text{-}in\text{-}\mathfrak{C}$), expressions for properties or relations of these objects ($\in$), variables for them ($x, \ldots$), quantifiers that bind these variables ($\forall x, \exists x, \ldots$), and rules of formation and inference, including logical rules for the quantifiers (such as, e.g., universal instantiation). Clearly, all of these conditions are satisfied here.[50] Indeed, Carnap's "variables of the new type" for the abstract objects introduced by a framework may be regarded as expressing in the formal mode what contemporary abstractionists would express in the material mode by: "abstraction may result in 'new' objects'..." (Linnebo, 2018, p.55). While Carnap did not make the abstraction process underlying the introduction of a new class of abstract objects by a framework explicit, the resulting abstract objects may certainly be qualified as *thin* "in the sense that their existence does not make a substantial demand on the world" (Linnebo, 2018, p.xi). Formulated less metaphysically, one might say that the concept of existence that is employed when the existence of logical objects of the framework is postulated within the framework is just that expressed by the purely logical $\exists x(Logical\text{-}in\text{-}\mathfrak{C}(x) \wedge \ldots)$, which is logically independent of the existence or non-existence of men, married people, bachelors, or other non-abstract objects.[51]

Analogously to the case of (R) before, the corresponding Ramsey sentence

(R_{Log}) $\exists R \exists S(\forall x(S(x) \rightarrow Logical\text{-}in\text{-}\mathfrak{C}(x)) \wedge ZF2[R, S])$

follows to be analytic(-in-$\mathfrak{C}$) if and only if there are R and S, such that $\forall x(S(x) \rightarrow Logical\text{-}in\text{-}\mathfrak{C}(x)) \wedge ZF2[R, S]$. Hence, if there are such R and S, then it will not just be the case that standard pure mathematics is analytic(-in-$\mathfrak{C}$) but additionally $\in$ and Set—as defined in $\mathfrak{C}$—will apply to objects that are $Logical\text{-}in\text{-}\mathfrak{C}$. If so, even 3d($Set$) and consequently 3c will be satisfied in $\mathfrak{C}$.[52]

The only downside would be that the corresponding question

(Q_{Log}) Are there R and S, such that $\forall x(S(x) \rightarrow Logical\text{-}in\text{-}\mathfrak{C}(x)) \wedge ZF2[R, S]$?

could no longer be addressed just by asking ordinary mathematicians or set theorists. For ordinary mathematicians are experts for ordinary mathematical objects and set theorists have additional expertise on sets, but neither are experts for logical objects, let alone logical objects in $\mathfrak{C}$. However, this remaining gap can be bridged: first add the set-theorists' terms Set and $\in$ to the vocabularly of the metalanguage of $\mathcal{L}$ in $\mathfrak{C}$; and then extend the deductive metalinguistic components of $\mathfrak{C}$ by the metalinguistic higher-order assumption that the logical relation $Val_{\mathfrak{M}}(\in)$ structures the logical objects in $Val_{\mathfrak{M}}(Set)$ in the same manner in which the set-theoretic membership relation $\in$ structures sets. That is:

(Ass_{Log}) $\langle Val_{\mathfrak{M}}(Set), Val_{\mathfrak{M}}(\in)\rangle \cong \langle Set, \in\rangle$.

With that in place, the previous high probability of there being R and S such that $ZF2[R, S]$, which resulted from set-theoretic considerations about $\langle Set, \in\rangle$, translates immediately into a high probability for there being R and S, such that $\forall x(S(x) \rightarrow Logical\text{-}in\text{-}\mathfrak{C}(x)) \wedge ZF2[R, S]$,

[50] Carnap (1950a, Section 3) actually speaks of the introduction of variables of a "new type", which would presuppose a many-sorted logic. But instead of introducing a new dedicated class of variables, one may just as well use one sort of variables and restrict them by the new general term $Logical\text{-}in\text{-}\mathfrak{C}$ instead.

[51] See Suppl. H of Leitgeb and Carus (2024) for more on Carnap on ontology.

[52] More should be said about what makes the members of the class $Logical\text{-}in\text{-}\mathfrak{C}$ properly *logical* (rather than just abstract). The key to this, in my view, would to argue that the members of $Logical\text{-}in\text{-}\mathfrak{C}$ might be regarded as abstract meaning-entities (Fregean senses or Carnapian intensions). But I will leave this to one side here.

which is the metalinguistic translation of (R_{Log}). Therefore, even R_{Log} ends up very likely analytic(-in-$\mathfrak{C}$). And the additional assumption (Ass_{Log}) hardly adds to the deductive strength of $\mathfrak{C}$, as it merely states that logicist sets and ordinary sets are structured alike, without saying which such sets exist and what their structure is like.

5. Conclusions

I have argued for (quasi-)Carnapian logicism: there is a logicist conceptual framework in which $\in$ and Set are defined in logical terms, and in which $ZF2[\in, Set]$ is (likely to be) semantically analytic. It follows that all standard terms of pure mathematics are logical in the framework, and all standard proven theorems of pure mathematics are (likely to be) semantically analytic in the framework. The required definitions, the semantic notion of analyticity in a framework, the logicist framework, and the occurrence and justification of the probabilistic qualification "likely to be" have been explained in the previous sections.

The essential features of the resulting Carnapian brand of logicism are: it is clear, formally precise, systematic, and reasonably simple. It still resembles mathematical practice in so far as it preserves the usual set-theoretic definitions of mathematical terms, it preserves the set-theoretic proofs of mathematical theorems, it acknowledges the open-endedness of the concepts of sethood and membership, it makes the existential presupposition of the set-theoretic treatment of mathematics explicit, and it incorporates (hypothetical) verdicts and arguments by ordinary mathematicians and set-theorists into its argument for the likely analyticity of that presupposition within the logicist framework. Its upshot is that pure mathematics can be rationally reconstructed as purely conceptual in the sense of coming along with a conceptual framework, while staying close to mathematical practice.[53] As shown in Section 4, the 'purely conceptual' can be extended even to the ontology of mathematics, to the effect that all mathematical objects are logical objects in the respective logicist framework. Finally, Carnapian Logicism is embedded in, and coheres with, Carnap's understanding of logic, theoretical terms, conceptual frameworks, analyticity, and probability and matches his overall conception of philosophy as rational reconstruction.

What Carnapian logicism does *not* achieve (and does not aim to achieve), as has been explained in Section 4, too, is to give mathematics a secure logical foundation. Epistemologically, it remains on the same level as Frege's *Grundlagen* in which Frege points out:

> I do not claim to have made the analytic character of arithmetical propositions more than probable. . . (Frege, 1884, *Die Grundlagen der Arithmetik*, §90)

In Frege's case, that was because the *Grundlagen* had not quite delivered sound, formally precise, and gap-free logical derivations of the mathematical laws of arithmetic from axioms of logic. That is what he hoped to supply in his later *Grundgesetze*, though we now know that he would fail to do so. In the meantime, Gödel's Incompleteness Theorems have made it seem unlikely that *any* logicist could do better than arguing for the analyticity of mathematics on probabilistic grounds.

[53] In contrast, empirical science could not be rationally reconstructed as purely conceptual while staying close to scientific practice. But it is not the place to argue for this claim here.

There might be one other potential downside to Carnapian Logicism: consider e.g. the Continuum Hypothesis, which we know is neither provable nor refutable in $ZF2[\in, Set]$. It is well-known that the Continuum Hypothesis can be reformulated as a statement CH in the language of pure second-order logic, and the same holds for its negation, which can also be expressed as a statement NCH in the same language (see Shapiro, 1991, p.105). Moreover, it is easy to see that either CH is logically true in full (model-theoretically defined) second-order logic or NCH is logically true in full second-order logic. For the same reason, either CH is analytic in the logicist framework from Section 3 or NCH is analytic in that framework, even though we do not know which of the two is the case. This matches Bohnert's (1975, p.211) summary of what Carnap told him in 1967, that is, "one could only wait and watch developments, with respect to what could be thought of as analytically true", where in the present case 'analytically true' would not be relative to our pre-theoretic understanding of set and membership but to the understanding afforded by the logicist framework from Section 3. At the same time, if it happened to be the case that mathematicians did not think CH is "settled" in that manner, this would amount to an important discrepancy between our logicist rational reconstruction of mathematics and what mathematicians would think themselves.[54] On the other hand, rational reconstructions are merely required to be similar to what they reconstruct; certain discrepancies are to be expected. And of course the deductive system of the logicist framework of this paper does not settle the question by means of proof, which might be all these mathematicians might mean by the Continuum Hypothesis not being settled. I will have to leave this question to future work.

In any case: as things stand, the resulting Carnapian logicist package does not seem to fare worse than any other philosophical interpretation of mathematics available. It is a coherent option that is on offer for anyone willing to choose it.

Acknowledgements. I am very grateful to an anonymous reviewer, Marianna Antonutti-Malfori, Johan van Benthem, Tim Button, André Carus, Cesare Cieslinsky, Fernando Ferreira, Salvatore Florio, Leon Horsten, Luca Incurvati, Dan Isaacson, Bruno Jacinto, Deborah Kant, Gregory Landini, Øystein Linnebo, Robert May, Colin McLarty, Julien Murzi, Uri Nodelman, Alex Paseau, Zeynep Soysal, Brett Topey, Pierre Wagner, Russell Wahl, Cheng Yong, and Ed Zalta for their help with earlier versions of this paper. This article is dedicated to Michael Friedman, who I am sure would have had numerous wonderfully enlightening comments on it. There is no one from whom I have learned more about Carnap than from Michael. He will be very much missed.

References

Avigad, J. and Zach, R. (2024). The Epsilon Calculus. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition. https://plato.stanford.edu/archives/fall2024/entries/epsilon-calculus/.

Awodey, S. and Carus, A. W. (2003). Carnap versus Gödel on Syntax and Tolerance. In Parrini, P., Salmon, W. C., and Salmon, M. H., editors, *Logical Empiricism: Historical and Contemporary Perspectives*, pages 57–64. University of Pittsburgh Press, Pittsburgh, PA. DOI: https://doi.org/10.2307/j.ctvt6rjh9.8.

[54] I am grateful to an anonymous reviewer for highlighting this.

Awodey, S. and Carus, A. W. (2004). How Carnap Could Have Replied to Gödel. In Awodey, S. and Klein, C., editors, *Carnap Brought Home: The View from Jena*, pages 203–223. Open Court, LaSalle, IL.

Awodey, S. and Klein, C., editors (2004). *Carnap Brought Home: The View from Jena*. Open Court, LaSalle, IL.

Boccuni, F. and Woods, J. (2020). Structuralist Neologicism. *Philosophia Mathematica*, 28(3):296–316. DOI: https://doi.org/10.1093/philmat/nky017.

Boghossian, P. A. (1996). Analyticity Reconsidered. *Nous*, 30(3):360–391. DOI: https://doi.org/10.2307/2216275.

Bohnert, H. G. (1975). Carnap's Logicism. In Hintikka, J., editor, *Rudolf Carnap, Logical Empiricist: Materials and Perspectives*, pages 183–216. Reidel, Dordrecht.

Carnap, R. (1929). *Abriß der Logistik, mit besonderer Berücksichtigung der Relationstheorie und ihrer Anwendungen*. Springer, Vienna.

Carnap, R. (1931). Die logizistische Grundlegung der Mathematik. *Erkenntnis*, 2:91–105. English translation: "The Logicist Foundations of Mathematics" in Benacerraf & Putnam (eds.), *Philosophy of Mathematics: Selected Readings* (1964), 41–52. DOI: https://doi.org/10.1007/BF02028142.

Carnap, R. (1934). *Logische Syntax der Sprache*. Springer, Vienna.

Carnap, R. (1937). *The Logical Syntax of Language*. Routledge, London.

Carnap, R. (1942). *Introduction to Semantics*. Harvard University Press, Cambridge, MA.

Carnap, R. (1946). Modalities and Quantification. *Journal of Symbolic Logic*, 11(2):33–64. DOI: https://doi.org/10.2307/2268610.

Carnap, R. (1950a). Empiricism, Semantics, and Ontology. *Revue Internationale de Philosophie*, 4(11):20–40.

Carnap, R. (1950b). *Logical Foundations of Probability*. University of Chicago Press, Chicago, IL.

Carnap, R. (1956). *Meaning and Necessity*. The University of Chicago Press, Chicago, IL. Second and enlarged edition; original 1947.

Carnap, R. (1959). Theoretical Concepts in Science. Unpublished manuscript; published in Stathis Psillos, "Rudolf Carnap's 'Theoretical Concepts in Science'", *Studies in History and Philosophy of Science Part A*, 31(1) (2000): 151–172.

Carnap, R. (1961). On the Use of Hilbert's ϵ-Operator in Scientific Theories. In Bar-Hillel, Y. et al., editors, *Essays on the Foundations of Mathematics*, pages 156–164. North-Holland, Amsterdam.

Carnap, R. (1963a). Intellectual Autobiography. In Schilpp, P. A., editor, *The Philosophy of Rudolf Carnap*, pages 3–84. Open Court, LaSalle, IL.

Carnap, R. (1963b). The Philosopher Replies. In Schilpp, P. A., editor, *The Philosophy of Rudolf Carnap*, pages 859–1013. Open Court, LaSalle, IL.

Carnap, R. (1966). *Philosophical Foundations of Physics: An Introduction to the Philosophy of Science*. Basic Books, New York, NY.

Carnap, R. (1971). A Basic System of Inductive Logic, Part I. In Carnap, R. and Jeffrey, R. C., editors, *Studies in Inductive Logic and Probability*, pages 33–165.

Carnap, R. and Jeffrey, R. C. (1971). *Studies in Inductive Logic and Probability, Vol. 1*. University of California Press, Berkeley, CA.

Demopoulos, W. (2007). Carnap on the Rational Reconstruction of Scientific Theories. In Friedman, M. and Creath, R., editors, *The Cambridge Companion to Carnap*, pages 248–272. Cambridge University Press, Cambridge, UK. DOI: https://doi.org/10.1017/CCOL9780521840156.012.

Džamonja, M. (2017). Set Theory and its Place in the Foundations of Mathematics: A New Look at an Old Question. *Journal of Indian Council of Philosophical Research*, 34(2):415–424. DOI: https://doi.org/10.1007/s40961-016-0082-6.

Ebbs, G. (2017). *Carnap, Quine, and Putnam on Methods of Inquiry*. Cambridge University Press, New York, NY. DOI: https://doi.org/10.1017/9781316823392.

Frege, G. (1884). *Die Grundlagen der Arithmetik: Eine logisch-mathematische Untersuchung über den Begriff der Zahl*. Wilhelm Koebner, Breslau.

Frege, G. (1893/1903). *Grundgesetze der Arithmetik, Vol. 1–2*. Hermann Pohle, Jena.

Friedman, M. (1999). *Reconsidering Logical Positivism*. Cambridge University Press, Cambridge, UK. DOI: https://doi.org/10.1017/CBO9781139173193.

Friedman, M. (2001). *Dynamics of Reason: The 1999 Kant Lectures at Stanford University*. CSLI Publications, Stanford, CA.

Gödel, K. (1995). Some Basic Theorems on the Foundations of Mathematics and Their Implications. In Feferman, S. e., editor, *Kurt Gödel Collected Works, Vol. 3: Unpublished Essays and Lectures*, pages 304–323. Oxford University Press, Oxford, UK. Originally a 1951 unpublished manuscript.

Hale, B. and Wright, C. (2001). *The Reason's Proper Study: Essays towards a Neo-Fregean Philosophy of Mathematics*. Clarendon Press, Oxford, UK. DOI: https://doi.org/10.1093/0198236395.001.0001.

Hilbert, D. (1899). Letter by Hilbert to Frege from December 29th 1899. In Gabriel, G., Hermes, H., Kambartel, F., Thiel, C., and Veraart, A., editors, *Gottlob Frege: Philosophical and Mathematical Correspondence*. Basil Blackwell, Oxford. 1980, pp. 38–43.

Hilbert, D. and Bernays, P. (1934/1939). *Grundlagen der Mathematik, Vol. I–II*. Springer, Berlin.

Hintikka, J. (1991). Carnap, the University of Language and Extremality Axioms. *Erkenntnis*, 35(1):325–336. DOI: https://doi.org/10.1007/BF00388292.

Hrbacek, K. and Jech, T. (1999). *Introduction to Set Theory*. Taylor & Francis, Boca Raton, FL, 3rd edition.

Incurvati, L. (2020). *Conceptions of Set and the Foundations of Mathematics*. Cambridge University Press, Cambridge, UK. DOI: https://doi.org/10.1017/9781108596961.

Kant, D. (2025a). *Pragmatic Insights into Set-theoretic Independence: Exploring Disagreement and Agreement among Practitioners*. Vittorio Klostermann, Frankfurt am Main.

Kant, D. (2025b). The Hidden Use of New Axioms. In Antos, C., Barton, N., and Venturi, G., editors, *The Palgrave Companion to The Philosophy of Set Theory*, pages 103–129. Palgrave Macmillan, Cham. DOI: https://doi.org/10.1007/978-3-031-62387-5_5.

Lavers, G. (2024). *Mathematics is (Mostly) Analytic*. Cambridge University Press, Cambridge, UK. DOI: https://doi.org/10.1017/9781009109925.

Leitgeb, H. (2021). On Non-Eliminative Structuralism. Unlabeled Graphs as a Case Study (Part B). *Philosophia Mathematica*, 29(1):64–87. DOI: https://doi.org/10.1093/philmat/nkaa009.

Leitgeb, H. (2023). Ramsification and Semantic Indeterminacy. *The Review of Symbolic Logic*, 16(3):900–950. DOI: https://doi.org/10.1017/S1755020321000599.

Leitgeb, H. and Carus, A. (2024). Rudolf Carnap. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition. https://plato.stanford.edu/archives/fall2025/entries/carnap/.

Leitgeb, H., Nodelman, U., and Zalta, E. N. (2025). A Defense of Logicism. *The Bulletin of Symbolic Logic*, 31(1):88–152. DOI: https://doi.org/10.1017/bsl.2024.28.

Lewis, D. K. (1970). How to Define Theoretical Terms. *The Journal of Philosophy*, 67(13):427–446. DOI: https://doi.org/10.2307/2023861.

Linnebo, Øystein. (2018). *Thin Objects*. Oxford University Press, Oxford, UK. DOI: https://doi.org/10.1093/oso/9780199641314.001.0001.

Linsky, B. and Zalta, E. N. (1994). In Defense of the Simplest Quantified Modal Logic. *Philosophical Perspectives*, 8:431–458. DOI: https://doi.org/10.2307/2214181.

Maddy, P. (2007). *Second Philosophy: A Naturalistic Method*. Oxford University Press, Oxford, UK. DOI: https://doi.org/10.1093/acprof:oso/9780199273669.001.0001.

Marschall, B. (2021). Carnap and Beth on the Limits of Tolerance. *Canadian Journal of Philosophy*, 51(4):282–300. DOI: https://doi.org/10.1017/can.2021.16.

Marschall, B. (2024). Herbert G. Bohnert: The Last Carnapian. *HOPOS*, 14(2):361–396. DOI: https://doi.org/10.1086/731679.

Parrini, P., Salmon, W. C., and Salmon, M. H. (2003). *Logical Empiricism: Historical and Contemporary Perspectives*. University of Pittsburgh Press, Pittsburgh, PA.

Paseau, A. and Pregel, F. (2023). Deductivism in the Philosophy of Mathematics. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition. https://plato.stanford.edu/archives/fall2023/entries/deductivism-mathematics/.

Pettigrew, R. (2008). Platonism and Aristotelianism in Mathematics. *Philosophia Mathematica*, 16(3):310–332. DOI: https://doi.org/10.1093/philmat/nkm035.

Quine, W. V. O. (1951). Two Dogmas of Empiricism. *The Philosophical Review*, 60(1):20–43. DOI: https://doi.org/10.2307/2181906.

Richardson, A. W. (1998). *Carnap's Construction of the World: The Aufbau and the Emergence of Logical Empiricism*. Cambridge University Press, Cambridge, UK. DOI: https://doi.org/10.1017/CBO9780511570810.

Russell, B. (1903). *Principles of Mathematics*. Cambridge University Press, Cambridge, UK.

Sattig, T. (2025). *How Time Passes*. Oxford University Press, Oxford, UK. DOI: https://doi.org/10.1093/9780198929192.001.0001.

Schiemer, G. (2013). Carnap's Early Semantics. *Erkenntnis*, 78(3):487–522. DOI: https://doi.org/10.1007/s10670-012-9365-8.

Schiemer, G. (2022). Logicism in Logical Empiricism. In Boccuni, F. and Sereni, A., editors, *Origins and Varieties of Logicism*, pages 243–266. Routledge, New York, NY.

Schiemer, G. and Gratzl, N. (2016). The epsilon-reconstruction of theories and scientific structuralism. *Erkenntnis*, 81:407–432. DOI: https://doi.org/10.1007/s10670-015-9747-9.

Shapiro, S. (1991). *Foundations without Foundationalism*. Clarendon Press, Oxford, UK. DOI: https://doi.org/10.1093/0198250290.001.0001.

Shapiro, S. (2008). Identity, Indiscernibility, and Ante Rem Structuralism: The Tale of i and $-i$. *Philosophia Mathematica*, 16(3):285–309. DOI: https://doi.org/10.1093/philmat/nkm042.

Shapiro, S. (2012). An 'i' for an '$-i$': Singular Terms, Uniqueness, and Reference. *Review of Symbolic Logic*, 5:380–415. DOI: https://doi.org/10.1017/S1755020311000347.

Skyrms, B. (2000). *Choice and Chance. An Introduction to Inductive Logic*. Wadsworth, Stamford. Fourth Edition.

Soysal, Z. (2025). The Problem of Existence for Descriptivism About the Reference of Set-Theoretic Expressions. In Antos, C., Barton, N., and Venturi, G., editors, *The Palgrave Companion to The Philosophy of Set Theory*, pages 11–36. Palgrave Macmillan, Cham. DOI: https://doi.org/10.1007/978-3-031-62387-5_2.

Sznajder, M. (2018). Inductive Logic as Explication: The Evolution of Carnap's Notion of Logical Probability. *The Monist*, 101(4):417–440. DOI: https://doi.org/10.1093/monist/ony015.

Tennant, N. (2013). Logicism and Neologicism. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2013 edition. https://plato.stanford.edu/archives/win2023/entries/logicism/.

Väänänen, J. and Wang, T. (2015). Internal Categoricity in Arithmetic and Set Theory. *Notre Dame Journal of Formal Logic*, 56(1):121–134. DOI: https://doi.org/10.1215/00294527-2835038.

Warren, J. (2020). *Shadows of Syntax: Revitalizing Logical and Mathematical Conventionalism*. Oxford University Press, Oxford, UK. DOI: https://doi.org/10.1093/oso/9780190086152.001.0001.

Whitehead, A. N. and Russell, B. (1910–1913). *Principia Mathematica, Vol. 1–3*. Cambridge University Press, Cambridge, UK.

Williamson, T. (1998). Bare Possibilia. *Erkenntnis*, 48(2):257–273. DOI: https://doi.org/10.1023/A:1005331819843.

Williamson, T. (2003). Everything. *Philosophical Perspectives*, 17(1):415–465. DOI: https://doi.org/10.1111/j.1520-8583.2003.00017.x.

Woods, J. (2014). Logical Indefinites. *Logique & Analyse*, (227):277–307.

Zermelo, E. (1930). Über Grenzzahlen und Mengenbereiche: Neue Untersuchungen über die Grundlagen der Mengenlehre. *Fundamenta Mathematicae*, 16:29–47.

Journal for the Philosophy of Mathematics

OPEN ACCESS

Citation: MEADOWS, Toby. (2025). The Consistency Hierarchy Thesis. *Journal for the Philosophy of Mathematics*. 2: 107-142. doi: 10.36253/jpm-2971

Received: September 18, 2024

Accepted: January 13, 2025

Published: December 30, 2025

ORCiD
TM: 0000-0003-2741-7685

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Firenze University Press
www.fupress.com

The Consistency Hierarchy Thesis

TOBY MEADOWS

Department of Logic and Philosophy of Science, University of California-Irvine, US.
Email: meadowst@uci.edu

Abstract: Set theorists often claim that natural theories are well-ordered by their consistency strength. We call this claim the *Consistency Hierarchy Thesis.* The goal of this paper is to unpack the philosophical and mathematical significance of this thesis; and to develop an understanding of how it is defended and, more particularly, how one might refute it. We shall see that the thesis involves a curious admixture of mathematics and philosophy that makes it difficult to pin down. We investigate some intriguing attempts to refute the thesis that are hampered by the problem of understanding what makes a theory natural. We then develop a thought experiment exploring the idea of what the ideal scenario for refutation would look like. And we show that a counterexample is impossible if we insist that the counterexample uses respectable (i.e., transitive) models. Finally, we reflect on how these hurdles affect our understanding of the significance of the thesis by drawing a parallel with a more famous claim: the Church-Turing thesis.

Keywords: Set theory, Forcing, Inner Model Theory, Consistency, Incompleteness.

He must,
so to speak,
throw away the ladder
after he has climbed up on it.

Wittgenstein

The following pair of facts are well-known. ZFC provides a practically adequate foundation for mathematics as we know it today; and yet, if ZFC is consistent, then it cannot be complete. As such, there are a dizzying variety of extensions of ZFC many of which are incompatible with each other. In the face of such chaos, one might be tempted to take up a conservative attitude and thus, prefer to remain within the comforting confines of ZFC. At present, such a move makes little difference to one's ability to found ordinary mathematics, but the curious mind will still wonder what is out there in the big beyond and just how wild the jungle is. Many set theorists have offered a tantalizing answer to the latter question: natural theories extending ZFC are well-ordered by their consistency strength. We call this the *Consistency Hierarchy Thesis*. The purpose of this paper is to explain what this thesis really means and to argue that it involves such a strange mix of mathematics and philosophy that it is difficult to know how one could successfully defend or rebut it.

We shall start in Section 1 by providing a gentle introduction to the consistency strength relation and the claim that it forms a hierarchy. While this material is very elementary, our patient discussion is intended to draw out the mathematical and philosophical agendas that drive our interest in the problem. In Section 2, we consider how one might attempt to refute the thesis and argue that the prospects for a successful refutation seem bleak. In particular, we shall focus on some proposed counterexamples from Joel David Hamkins' recent paper on the topic (Hamkins, 2025). While we shall push back on these proposals, the emerging theme will not be so much that Hamkins is wrong, so much as that there is something odd about the question itself. Finally, in Section 3, we'll offer some explanation as to why the thesis is so difficult to rebut and reflect on how this should impact our understanding of it.

1. What is the thesis and why is it important?

1.1. What is relative consistency?

Let T be a theory in the language of set theory, $\mathcal{L}_{\in}$. Recall that a theory T is consistent if we cannot prove both φ and $\neg\varphi$ using assumptions from T. Using some form of Gödel coding we may formulate a statement, $Con(T)$, in the language of arithmetic that says T is consistent; or more formally, we have

$$T \text{ is consistent } \Leftrightarrow \mathbb{N} \models Con(T).$$

Further, if T can interpret PA as ZFC does, then such a statement can be reasonably formulated in $\mathcal{L}_{\in}$. Thus, we might naturally ask whether T can prove its own consistency. Of course, this was scotched by Gödel.

Proposition 1. *(Gödel)* $PA \nvdash Con(PA)$ *if* PA *is consistent.*

A quick perusal of the proof reveals that it continues to hold for any theory T that can interpret PA. It also gives us our first example of a relative consistency proof. It tells that if PA is consistent, then $PA + \neg Con(PA)$ is also consistent. We might say that $PA + \neg Con(PA)$ is *consistent relative* to PA.

That's the basic idea of relative consistency, but we need to take a little more care if we are to formulate an interesting mathematical relation. To illustrate this, note that we *do* think PA is consistent, and so we think that PA cannot prove $Con(PA)$. Thus, it seems reasonable then to think that $PA + Con(PA)$ should be stronger than PA and so, $PA + Con(PA)$ should not be consistent relative to PA. But as we've defined things so far, this is false. To see this note that the statement,

$$Con(PA) \rightarrow Con(PA + Con(PA))$$

is true simply because the consequent is true.[1] The upshot is that if we work in a background theory like ZFC, then every pair of theories for which ZFC can provide a model will be consistent relative to each other; i.e., equiconsistent. Or as a slogan: every pair of consistent theories would be equiconsistent with each other. Thus, the relative consistency relation is rendered trivial for all theories weaker than the background theory we are working in. This would be an uninteresting mathematical relation. The moral of this story is that we need to pay

[1] More specifically, assume standard mathematical conventions and work in ZFC. Then we can define a standard model $\mathbb{N}$ of arithmetic based on ω that satisfies PA. By soundness, this implies that $Con(PA)$ is true. Moreover, since $\mathbb{N}$ is standard it agrees with the universe on all arithmetic statements. Thus, $\mathbb{N} \models PA + Con(PA)$ as required.

attention to where we are standing when we prove these theorems. In particular, if we weaken our background theory to PA, then the statement above is no longer provable there. Or more formally,

$$PA \nvdash Con(PA) \rightarrow Con(PA + Con(PA))$$

since if it were provable an application of the deduction theorem would violate Gödel's second theorem. Thus, if we want an interesting mathematical relation we need to ensure that our background theory is weak enough not to trivialize it.

It's also worth noting the degree of metamathematics in the statement above. We aren't simply providing a counterexample to a conditional. To establish a failure of relative consistency, we need to prove that there is no proof in PA that if there is no proof of absurdity from PA, then there is no proof of absurdity from PA plus the statement that there is no proof of absurdity from PA. Obviously, the formal notation employed above makes it easier to articulate such statements, but we are – I think – far enough down the Gödelian rabbit hole that analyzing their philosophical significance becomes challenging. While it is certainly possible to reason accurately in these domains, it is not so obvious that our naive intuitions about provability come along for the ride.

In this paper, we shall be predominantly concerned with theories that extend ZFC. As such, we'll still have an interesting mathematical relation if we let ZFC be our background theory. With this in mind, we then offer the following definition:

Definition 2. For theories T and S extending ZFC in the language of set theory, let us say that T is *consistent relative to* S, abbreviated $T \leq_{Con} S$ if[2]

$$ZFC \vdash Con(S) \rightarrow Con(T).$$

If $T \leq_{Con} S$ and $S \leq_{Con} T$ we say that S and T are *equiconsistent*, abbreviated $\equiv_{Con}$. If $T \leq_{Con} S$ but $S \not\leq_{Con} T$, let us say that T is *properly consistent relative to* S, *abbreviated* $T <_{Con} S$.

We then observe that we have:

$$ZFC + \neg Con(ZFC) \equiv_{Con} ZFC <_{Con} ZFC + Con(ZFC).$$

Thus, we have a relation on theories that isn't trivial above ZFC. This raises a natural mathematical question: what kind of relation is $\leq_{Con}$? This is the kind of question a mathematician can get stuck into.

1.2. What is the thesis?

For the last sixty years, set theorists have been developing a better understanding of $\leq_{Con}$. The seminal results are from Gödel and Cohen.

Theorem 3. *(1, Gödel, 1940)* $ZFC + CH \leq_{Con} ZFC$*; and*
(2, Cohen, 1963) $ZFC + \neg CH \leq_{Con} ZFC$.

The first result is obtained by defining an *inner model* L of the universe in which $ZFC + CH$ holds. The second is obtained by taking a model of ZFC and using *forcing* to *generically*

[2]It's worth noting that little would change if we'd use PA as a base theory as almost all examples of relative consistency proofs in set theory can be carried out there. See Chapter VII.9 of (Kunen, 2006) for more discussion. One drawback with using PA is that we don't have the soundness and completeness theorems available, which tends to make proofs longer and more tedious.

extend and thus, obtain a model where ZFC is preserved but CH fails. Since the advent of these results, the techniques of inner model theory and forcing have developed substantially and become two of the mainstays of contemporary set theory. This work has lead to a much clearer understanding of the $\leq_{Con}$ relation and an intriguing answer to our question above. Moreover, this answer brings us to the headline of this paper: the *Consistency Hierarchy Thesis*. For a working version, we turn to John Steel:[3]

> If T is a natural extension of ZFC, then there is an extension H axiomatized by large cardinal hypothesis such that $T \equiv_{Con} H$. Moreover, $\leq_{Con}$ is a prewellorder of the natural extensions of ZFC. In particular, if T and U are natural extensions of ZFC, then either $T \leq_{Con} U$ or $U \leq_{Con} T$. (Steel, 2014)

Here we see the intriguing claim from the introduction of this paper. Instead of chaos, it is claimed that we have order. For ease of reference and uniformity of notation, let's break the quote above into its separate claims.

(CHT1) For all natural theories T extending ZFC, there is some extension LC_T of ZFC by a large cardinal axiom such that

$$T \equiv_{Con} LC_T;$$

(CHT2) $\leq_{Con}$ is a prewellordering on natural theories extending ZFC; and
(CHT3) For any pair S, T of natural theories extending ZFC either $T \leq_{Con} S$ or $T \leq_{Con} S$.

Clearly (CH3) follows from (CH2) and later we'll see that there is also a sense in which (CH2) follows from (CH1). Together, we'll call them the *Consistency Hierarchy Thesis* (CHT). Mathematically speaking, this is a surprising and interesting claim. (CH1) tells us that every natural theory is aligned with a large cardinal axiom. (CH2) then tells us that these theories are ordered in about as clean a way as one could want. Beyond being a pleasing arrangement, it also suggests that a serious mathematical idea is being chased, if it is true. But it is important to understand that CHT is not a theorem. The problem is not so much that we don't know whether it's true or not. The problem is just not stated with sufficient precision to even be amenable to proof or refutation.[4] This the first place where we see something extra-mathematical or even philosophical creeping into our discussion. The problem is that two of the terms used in CHT's formulation lack precise definitions: large cardinals; and natural theories.

When we speak of large cardinals, we generally think of them implying the existence of an elementary embedding from the universe into an inner model with certain closure properties.[5] For example, there is a measurable cardinal iff there is an elementary embedding $j : V \to M$ such that j moves at least one ordinal, the least of which is known as its critical point. Despite the availability of workable rules of thumb, there is – at least at present – no definition of large cardinal that captures all known large cardinals in a satisfying way that also leaves room for the future.[6] Nonetheless, we do currently have an enumeration of a large collection of large cardinal

[3] I should note that Steel calls this the *vague conjecture*.
[4] By proof here, I have in mind the kind of proofs that are written by mathematicians. One might also think that philosophers can deliver philosophical proofs, although we shall avoid that usage in this paper. This is pertinent to our discussion of Church's thesis below. An excellent article, which takes a different stance to that in this paper can be found in (Black, 2000).
[5] I'm ignoring smaller large cardinals like inaccessible and Mahlo cardinals here.
[6] Sometimes definitions of restricted class of large cardinal are useful. See for example (Woodin, 2001).

axioms that appear to be sufficient for at least our current purposes and which we don't know how to extend in a meaningful way.[7]

Natural theories, on the other hand, are a larger thorn in our side. They play a crucial role in CHT and they also lack a precise mathematical definition. In his discussion of CHT, Steel offers the following informal characterization:

> By "natural" we mean considered by set theorists, because they had some set-theoretic idea behind them. Here the standards are very liberal, as the many thousands of pages published by set theorists will testify. (Steel, 2014)

The general idea here is that a theory extending ZFC is *natural* if it is the sort of thing a set theorist might come up with when investigating some mathematical project. This is quite vague and indeed, there will certainly be problems at the borderline. Nonetheless, it's not difficult to identify some prototypical *in* and *out* cases.

On the *in*side, we have large cardinal axioms, forcing axioms, determinacy axioms, ultrapower axioms, generalized large cardinal axioms and perhaps dilator axioms (Goldberg, 2022; Kanamori, 2003; Lewis, 1998; Martin, nd; Todorcevic, 2014). Each of these is associated with a project that generalizes a mathematical question beyond the reach of ZFC and searches for axioms that address the problems in the expected way. For a classic example, it is well-known that Vitali sets provide examples of sets of reals that are not measurable. This raises questions about how and where this apparent pathology emerges. In the context of ZFC, the classical descriptive set theorist, Luzin, was able to show that every $\boldsymbol{\Sigma^1_1}$ set is Lebesgue measurable[8], but further progress was hampered by the limitations of ZFC. The use of determinacy axioms proved fruitful here. For example, Kechris and Martin showed that if every game on a $\boldsymbol{\Sigma^1_n}$ set of reals is determined, then every $\boldsymbol{\Sigma^1_{n+1}}$ set is Lebesgue measurable [9]. Thus, by extending ZFC with determinacy axioms a larger family of sets of reals could be tamed. Moreover, the addition of determinacy axioms seems to provide a very natural generalization beyond ZFC of the work carried out by classical descriptive set theorists in the mid twentieth century.[10] We might say that determinacy axioms provide archetypal examples of natural theories extending ZFC.

On the *out*side, typical examples of unnatural theories tend to involve what one might think of as metamathematical as opposed to combinatorial content. The axioms involved often make essential use of coding tricks and self-reference. For a classic example, we might consider the theory obtained by adding the statement $\neg Con(ZFC)$ to ZFC. This gives us a theory that talks about itself using coding and makes the bizarre claim that it is itself inconsistent. Beyond being a very odd thing to say, one might think of it as an unlikely thing for a mathematician to come up with when thinking about and working in set theory. There is a sense in which this theory doesn't talk about sets, but rather about the theory of sets. Riffing on Quine we might think that it is *mentioning* ZFC rather than *using* it. As such, we might think that it fails to satisfy Steel's criterion above. I should, however, say that I am quite skeptical about how robust the distinction between metamathematical and combinatorial content is. While anyone who has

[7] Of course, one can always take a large cardinal and move its successor. So roughly speaking by "meaningful" I mean extending by some means that we haven't already used in the large cardinals that come before it.

[8] See Theorem 29.7 in (Kechris, 1995).

[9] See Exercise 27.14 of (Kanamori, 2003).

[10] For a good discussion of the generalization of classical descriptive set theory see (Maddy, 2011) and (Martin, 1998).

thought about Gödel's incompleteness theorems will be aware that coding and self-reference seem to shift our focus away from the ordinary content of a theory, such content was always already there in the math. We just didn't see it until relatively recently. Moreover, anyone who has tried to understand how the logic of consistency statements works, will know that a new layer of combinatorial content emerges at this metamathematical level in the form of something like a modal logic.[11] Thus in this paper, we shall not take it that the use of metamathematical tools as sufficient for the identification of an unnatural theory.[12]

Despite these reservations, I still think it is often relatively easy to distinguish clear cases of metamathematical and combinatorial content.

Because of these ambiguities, our reasons for thinking that CHT is correct are based on what is essentially empirical evidence. Most of the natural theories that we have happened upon so far have been shown to be equiconsistent with a large cardinal extensions of ZFC. While many open questions remain, there are – as yet – no accepted counterexamples to CHT. As such, CHT has something in common with another famous thesis: the Church-Turing thesis. In that case, a major reason for thinking that it is true is that *all* of the models of informal computation we have come across so far have been shown to be equivalent to Turing machines. But there are also some important differences. While like the Church-Turing thesis, CHT has no accepted counterexamples, unlike the Church-Turing thesis, many equivalences remain unproven. For example, while the very educated guess is that the addition of a supercompact cardinal is equiconsistent with the addition of the proper forcing axiom to ZFC, this problem has remained open for over fifty years. So in contrast to the Church-Turing thesis, the picture with CHT is very far from complete, although there seems to be no compelling reason for pessimism.

All of this puts CHT in a remarkable position. I think it's clearly a claim that warrants mathematical attention. The $\leq_{Con}$ ordering is non-trivial and there is structure to be investigated. Indeed, the study of the $\leq_{Con}$ ordering has frequently provided some of the deepest results in set theory and inspired the development of new techniques. But CHT also makes use of extra-mathematical content in the form of large cardinals and natural theories. This certainly makes for an intriguing collection of set theoretic problems. But beyond this, is there any philosophical content to CHT? Or is this just another curio in a minority sport?

1.3. Why is it important?

Now that we've explained why CHT is a mathematically interesting problem, I want to shift our focus to its philosophical and foundational significance. I aim to demonstrate that CHT exerts a weighty influence on our understanding of set theory today. In particular, I will argue that: CHT provides a kind of evidence for the importance of large cardinals axioms; and CHT provides an explanation for the lack of disagreement between strong set theories with regard to concrete mathematical questions.

1.3.1. The importance of large cardinals

Recall that (CHT1) says that every natural theory is equiconsistent with a large cardinal extension of ZFC and that (CHT2) says $\leq_{Con}$ pre-well-orders the natural theories. The latter

[11] See, for example, (Boolos, 1984).

[12] This means that the discussion of putative counterexamples to CHT in Section 2 will need to be quite detailed and often formal. We shall need to demonstrate *why* certain metamathematical techniques give us unnatural theories.

of these is particularly significant since it would tell us that natural theories are ordered in a extremely tidy manner. But more than this, it would establish that our initial worries about chaos in the array of theories beyond ZFC were premature: we would have a much tamer realm than we could have reasonably hoped for. I take it that this insight is of philosophical and foundational importance for set theory and both mathematics and logic more generally. I would now like to demonstrate that there is a sense in which (CHT2) follows from (CHT1). Thus, we shall see that the purported tameness of natural theories is crucially dependent on our use of large cardinal axioms.

To see this, we start by observing that – unlike natural theories, in general – large cardinal extensions of ZFC tend to be very easily ordered by consistency strength. To illustrate this, we describe three grades of ordering that often occur in these relationships.[13] First, we note that very often a large cardinal that is stronger (in terms of consistency strength) than another will already satisfy the characteristic property of the weaker large cardinal. For example, measurable cardinals are stronger than inaccessible cardinals, and whenever κ is a measurable cardinal, κ is inaccessible. Thus, the consistency of ZFC plus a measurable cardinals implies the consistency of ZFC and an inaccessible cardinal. Second, we observe that even if the first condition doesn't hold, there will often be an example of the weaker large cardinal below the stronger one. For example, Woodin cardinals are stronger than measurable cardinals, but Woodin cardinals are not typically measurable. Nonetheless, there will be many measurable cardinals below a Woodin cardinal. This also gives the desired relative consistency fact. Third and finally, even when the former two situations don't occur, we may almost always use the stronger large cardinal to very easily define a model[14] where the weaker large cardinal axiom is satisfied. For example, Woodin cardinals are stronger than strong cardinals, but generally, there are no strong cardinals below a Woodin cardinal. Nonetheless, whenever δ is a Woodin cardinal, V_δ will think there are many strong cardinals. Thus again, we obtain the desired relative consistency claim.

There are, of course, examples that don't fit into any of the three grades, but such exceptions are comparatively rare.[15] The point that we are trying to emphasize is that the determination of consistency strength relationships between large cardinals is quite elementary. Moreover, it is easy to see that theories extending ZFC whose relative consistency can be ascertained by one of the three grades above will be pre-well-ordered by consistency strength. Thus, while there is some vagueness in the concept of a large cardinals, we can have much greater confidence that the strength of large cardinal axioms are pre-well-ordered, in contrast to the more difficult question regarding arbitrary natural theories. Now if we combine this observation with (CHT1), we see that (CHT2) follows immediately. If every natural theory is equiconsistent with a large cardinal extension of ZFC, then clearly the consistency strength of natural theories is also pre-well-ordered.

Thus, we see that large cardinals play a crucial supporting role in CHT, but in the – so to speak – other direction, CHT and the considerations above also tell us that large cardinals are a very special kind of axiom. They provide the underlying spine that brings order to the chaotic world opened by Gödel almost one hundred years ago. But before we fall prey to delusions of grandeur, I'd prefer to avoid pushing on and arguing that this somehow gives us evidence that

[13] We shall provide a more detailed sketch of one of these arguments in Section 3.2.1., but for now we'll content ourselves with a little vagueness.

[14] In particular, the model will be a rank initial segment of the universe satisfying the strong axiom.

[15] An obvious example occurs, if we admit that saying $0^\#$ exists is a large cardinal axiom. Then $0^\#$ is stronger than an inaccessible cardinal, but there are models where $0^\#$ exists that have no rank initial segment with an inaccessible cardinal. There are of course inner models, in particular, L that will have inaccessible cardinals, but I'm excluding this from the template since arguably L is a more sophisticated internal model to define than a rank initial segment.

large cardinal axioms are true. I think any attempt to answer that question would open a can of philosophical worms outside our interests here. But more importantly, I think addressing such questions could detract from what I take to be the surefooted lesson we can take from these observations. If one is interested in strong mathematical theories, then to the best of our current knowledge, large cardinal axioms are an essential tool in this investigation.

1.3.2. The absence of disagreement

For our next illustration of the foundational significance of CHT, we turn to the phenomenon of disagreement, or rather the lack thereof between strong natural theories regarding concrete mathematics. We've seen above that CHT provides order to the variety of theories extending ZFC, but we might still worry that the remaining options might affect ordinary mathematics regardless of whether these options are equiconsistent with a large cardinal axiom. For example, we know from Gödel that the theory of analysis is incomplete. And we know that ZFC and its extensions can fill in some of the gaps left by this incompleteness. Thus, there is a natural worry that the different theories extending ZFC will give different answers to problems in ordinary mathematics, in particular questions about the real numbers. It turns out that this is not the case. To explain this, we first need a refinement of (CHT1). In actual practice to date, we not only find that natural theories are equiconsistent with large cardinal extensions of ZFC, but that there is also a certain uniformity in the techniques used for the proofs of these facts. In particular, we generally find that:

(GE) Models of natural theories are obtained through *generic extension* of models of large cardinal extensions of ZFC; and
(IM) Models of natural theories deliver *inner models* of large cardinal extensions of ZFC.

This is a much stronger relationship than mere equiconsistency. For example, the models obtained by either generic extension or inner model theory retain exactly the same ordinals as the model we started with. This kind of structural preservation leads to theories that must agree with each other on substantial fragments of ordinary mathematics. For example, it can be argued that any pair of natural theories extending ZFC must agree on all Π^1_2 statements.[16] Moreover, as we consider strengthenings of ZFC by large cardinals, agreement on more mathematics can be obtained. Rather than describe a specific example of this phenomenon, we shall offer a general sketch of the proof strategy with the aim of highlighting the role played by CHT.

Theorem 4. *(Sketch[17]) Let T and S be natural theories extending ZFC; and let Γ be a fragment of the theory of analysis. Now suppose that there are large cardinal extensions LC_T and LC_S of ZFC, which are comparable by one of the three grades discussed above, such that:*

(i) LC_T interprets T via forcing that preserves Γ;[18]
(ii) T interprets LC_T via an inner model interpretation that preserves Γ;[19]
(iii) LC_S interprets S via forcing that preserves Γ; and

[16] See Steel's article in (Feferman et al., 2000; Maddy and Meadows, 2020; Meadows, 2021; Steel, 2014) for more detailed discussion of this.
[17] I've called this "theorem" a "sketch" since it is not stated with proper mathematical precision. Like the statement of CHT, it makes use of the imprecise terms: natural theory and large cardinal. Nonetheless, it seems it seems beneficial to draw out and highlight the underlying strategy in these arguments.
[18] By this we mean that LC_T can prove there is a poset $\mathbb{P}$ such that: $\Vdash_{\mathbb{P}} T$; and for all $\varphi \in \Gamma$, $\varphi \leftrightarrow \Vdash_{\mathbb{P}} \varphi$.
[19] By this we mean that T can prove there is a definable inner model N such that: $(LC_T)^N$; and for all $\varphi \in \Gamma$, $\varphi \leftrightarrow \varphi^N$.

(iv) S interprets LC_S via an inner model interpretation that preserves Γ.

Then T and S agree upon Γ; i.e., there is no sentence $\varphi \in \Gamma$ such that $T \vdash \varphi$ and $S \vdash \neg\varphi$, and there is no sentence $\varphi \in \Gamma$ such that $S \vdash \varphi$ while $T \vdash \neg\varphi$.[20]

Proof. (*Sketch*) Using CHT, we know that $LC_S \leq_{Con} LC_T$ or $LC_T \leq_{Con} LC_S$. We first the former and thus, that LC_T is stronger than LC_S. In this case, it will suffice to show that for all $\varphi \in \Gamma$ if $S \vdash \varphi$, then $T \vdash \varphi$ also. We proceed by contraposition and let $\mathcal{M}$ be a countable model satisfying $T \cup \{\neg\varphi\}$. Then let $N^{\mathcal{M}}$ be an inner model of $\mathcal{M}$ given by (2), so we $N^{\mathcal{M}}$ is a model of $LC_T \cup \{\neg\varphi\}$. Now since LC_T is stronger than LC_S as witnessed by one of our three grades of comparison, we may fix a model $V_\alpha^{N^{\mathcal{M}}}$ inside $N^{\mathcal{M}}$ satisfying LC_S.[21] Moreover, since $V_\alpha^{N^{\mathcal{M}}}$ is a rank initial segment of $N^{\mathcal{M}}$, they share the same theory of the reals, so $V_\alpha^{N^{\mathcal{M}}}$ also satisfies $\neg\varphi$. Finally, we use (3) to obtain a generic extension $V_\alpha^{N^{\mathcal{M}}}[G]$ of $V_\alpha^{N^{\mathcal{M}}}$ that satisfies $S \cup \{\neg\varphi\}$. Thus, we've established that $S \nvdash \varphi$. A similar argument works in the second case. □

For an example of this, let T be ZFC plus Δ^1_2-determinacy; and let S be ZFC plus the existence of a precipitous ideal. Then it can be shown via inner models and forcing that T is equiconsistent with LC_T where LC_T is the extension of ZFC with a Woodin cardinal.[22] By similar means it can be shown that S is equiconsistent with LC_S where LC_S is the extension of ZFC with a measurable cardinal. Moreover, it can be seen that these interpretations satisfy conditions (1) through (4) for Π^1_3 statements.[23] Finally, it is not difficult to see that $LC_S <_{Con} LC_T$ and so the argument strategy above tells us that T and S agree on Π^1_3 statements. The hard work in proving examples of this theorem sketch goes into establishing that the theories in question are sufficient to deliver conditions (1) to (4) for Γ. But the underlying structure of the proof is relatively simple: we move from models of natural theories to inner models of large cardinals in the spine and then slide down and generically extend outward to a model of another natural theory.

Thus, in addition to establishing the value of large cardinals as an instrument for understanding strong set theories, we see that CHT also provides a valuable explanation of why disagreement about concrete mathematics does not seem to occur between strong natural theories. Thus, we've now demonstrated that beyond being an interesting mathematical problem, CHT has deep philosophical implications for set theory and the way in which it provides a foundation for mathematics. Hopefully, it is clear that if a counterexample to CHT were discovered and accepted, then set theory as we know it would undergo a radical shift in perspective.

2. Attempts at refutation

Now that we have a clearer idea of what the Consistency Hierarchy Thesis is and why it is foundationally significant, we are going to shift our attention to the question of whether it is justified. We saw above that our reasons for believing it rest upon the fact that we have found many extensions of ZFC that satisfy it, and the fact that no counterexamples to it have been discovered and accepted by the set theory community. We likened CHT to the Church-Turing

[20] It might be more apropos to say that *S and T don't disagree on Γ*, but I've opted for the simpler slogan to avoid using too much negation.
[21] Here I'm abusing notation a little and allowing that α could be $Ord^{\mathcal{M}}$ for the case of the first grade.
[22] See Theorem 32.17 in (Kanamori, 2003).
[23] See Theorem 15.6 in (Kanamori, 2003).

thesis, but noted that in contrast the picture with CHT is much less complete. While many equivalences with large cardinal axioms have been established many also remain open. As such, a much wider flank is left exposed to the possibility of a counterexample. In this section, my goal is to consider the prospects for such a campaign. We shall begin by considering a counterexample that almost nobody thinks is significant. Then we shall consider a series of more serious attempts from a recent paper by Joel David Hamkins (2025). While we shall push back on Hamkins' intriguing examples, the real value of this work will be in illuminating more about just what kind of problem CHT presents.

2.1. A counterexample that nobody accepts

In this section, we give a short proof that there is a pair of theories whose consistency strengths are incomparable and consider what this means for CHT. The example is taken from Theorem 3 of (Hamkins, 2025), but it will benefit us to provide a quick sketch in order emphasize the technical nature of the reasoning involved and to illustrate how little purchase our ordinary intuitions about proof have in this arena. In particular, this will help us get a better understanding of why people are not impressed by it as a putative counterexample to CHT. Rather than providing a detailed discussion of notational conventions, I'll just note that I'm aiming to follow standard conventions as one might find in (Lindström, 2003).[24] Recall that if η is a Rosser sentence for the theory T, then η says that if d is (code of) a T-proof of η, then there is some $e < d$ that is a T-proof of $\neg\eta$.

Theorem 5. *Let η be the Rosser sentence for $ZFC + Con(ZFC)$ and let T denote this theory. Then supposing that $ZFC + Con(ZFC)$ is consistent, the consistency strengths of $Con(ZFC + \eta)$ and $Con(ZFC + \neg\eta)$ are incomparable; i.e.,*

(i) $ZFC \nvdash Con(ZFC + \neg\eta) \rightarrow Con(ZFC + \eta)$;
(ii) $ZFC \nvdash Con(ZFC + \eta) \rightarrow Con(ZFC + \neg\eta)$.

Proof. (*Sketch*) (1) Our plan is to show there is a model $\mathcal{M}$ of ZFC satisfying:

(i) $Con(ZFC + \neg\eta)$; and (ii) $\neg Con(ZFC + \eta)$.

For (i) we note since η is a Rosser sentence for T, we may fix a model $\mathcal{M}$ satisfying $T + \neg\eta$. Then unpacking the definition of $\neg\eta$, we see that $\mathcal{M}$ thinks there is a T-proof d of η and there is $e < d$ where e is a T-proof of $\neg\eta$. It can be seen that this statement is Σ^0_1 and so since ZFC is Σ^0_1-complete, we see that $\mathcal{M}$ thinks ZFC can prove it. Thus, $\mathcal{M}$ thinks that $ZFC \vdash \neg\eta$; i.e., $\mathcal{M} \models \neg Con(ZFC + \eta)$. For (ii), we simply observe that in $\mathcal{M}$ we have both $ZFC \nvdash \bot$ and $ZFC \vdash \neg\eta$ and so $ZFC \nvdash \eta$. Thus, $\mathcal{M} \models Con(ZFC + \neg\eta)$ as required.

(2) We show there is a model $\mathcal{M}$ of ZFC satisfying:

(i) $Con(ZFC + \eta)$; and (ii) $\neg Con(ZFC + \neg\eta)$.

[24] Of course, that book is focused on arithmetical theories rather than set theories but this makes little difference since ZFC is an essentially reflexive theory that interprets PA. I'll also note that when context requires it below, we shall have recourse to be more precise about our notational conventions.

For (i), we start by using Gödel's second theorem to obtain a model $\mathcal{M}$ satisfying:

$$[T + \eta] + \neg Con([T + \eta]).$$

Then with a little Rosser-style reasoning, it can be seen that $\mathcal{M}$ satisfies the statement: "there is a T-proof e of $\neg\eta$ and for all $d < e$, d is not a T-proof of η." And since this statement is Σ^0_1, we see that ZFC is able to prove it and so $\mathcal{M}$ thinks ZFC proves it too. With a little more work, it can be seen that ZFC proves that the statement implies η and so $\mathcal{M}$ thinks that $ZFC \vdash \eta$; i.e., $\mathcal{M} \models Con(ZFC + \neg\eta)$. For (ii), we now know $\mathcal{M}$ thinks that $ZFC \nvdash \bot$ and $ZFC \vdash \eta$. Thus, $\mathcal{M} \models Con(ZFC + \eta)$ as required. □

Thus, we see that the theories $ZFC + \eta$ and $ZFC + \neg\eta$ have consistency strengths that are incomparable. So we see very clearly that $\leq_{Con}$ is not a pre-well-ordering on all theories. What should we make of this? Why shouldn't we think of this as being a counterexample to CHT? I suppose for some people it might be, but the general consensus is that this counterexample does not succeed. The obvious culprit is the unnaturalness of the theories involved. But why should we think that $ZFC + \eta$ and $ZFC + \neg\eta$ are unnatural? First recall that if we put an intuitive gloss on things, η says something like

> If I'm provable, then I have already been refuted.

It's quite an odd sentence. It's conjured using coding and the diagonal lemma to achieve the effect of self-reference. Moreover, it's relatively obvious that the resulting theories were cooked up for the specific purpose of delivering a counterexample. We suggested above that these were indicators of unnaturalness and I think it's probably fair to say that the tools used above are not generally elements of the ordinary mathematician's toolkit. It's not the sort of theory that one would expect a mathematician doing set theory to come up while they were working with set theory when it is understood as a theory of sets. As such, I think we have reason to think that they do not satisfy Steel's criterion for being a natural theory.

But we can also make a more specific complaint in this case. Assuming we are comfortable accepting some large cardinals, it is easy to see that η is simply true.[25] But what do we learn from this? First, note that since any ω-model of ZFC will satisfy $Con(ZFC)$, we see that if η is false in a model $\mathcal{M}$ of ZFC then $\mathcal{M}$ cannot be an ω-model. I think it would be very strange to accept a theory extending ZFC that cannot be satisfied in a model with the genuine natural numbers. We'd have good reason to think that such a theory is defective. But more than this, it is also easy to see that $ZFC + \neg\eta$ is ω-inconsistent.[26] As is well-known, Gödel's original proof of the incompleteness theorem merely assumed that his target theory was ω-consistent rather than simply consistent (Gödel, 1986). While the result was later improved by Rosser, the assumption of ω-consistency was thought to be sufficient for the philosophical impact of his theorem to be felt. As such, it seems reasonable to exclude $ZFC + \neg\eta$ from the realm of natural theories.

Hopefully, this example illustrates the magnitude of the problem of refuting CHT. It is not sufficient to provide a pair of theories whose consistency strengths cannot be be compared. We must also show that those theories are natural. In the absence of any precise

[25] Very briefly, suppose there is an inaccessible cardinal κ and suppose toward a contradiction that η is false. Then there is a $(ZFC + Con(ZFC))$-proof of η and so η is true in every model of $ZFC + Con(ZFC)$. V_κ is such a model, so η would be true there, and since V_κ and the universe shares their natural numbers, η would be true, which contradicts our initial assumption.

[26] It is easy to see that $ZFC + \neg\eta$ must prove that there is some d that is a $(ZFC + Con(ZFC))$-proof of η; and yet for all $n \in \omega$, $ZFC + \neg\eta$ will also prove that n is not a $(ZFC + Con(ZFC))$-proof of η.

definition, a significant part of the counterexample challenge is philosophical, or at least, extra-mathematical. In particular, we need to mount an argument, as opposed to a proof, establishing that the theories in question are natural. This makes for quite a strange mathematical problem, and one that might not be amenable to a definitive resolution.

2.2. Some counterexamples that probably don't succeed

Our goal now is to consider some serious attempts to answer this refutation challenge. For this, we turn to Hamkins' recent paper on the topic (Hamkins, 2025). Among other things, this paper provides a helpful and elegant overview of literature regarding CHT. Hamkins then proposes a number of putative counterexamples to CHT including an impressive array of generalizations of the techniques involved.[27] Given the discussion of the previous section, we shall be most interested in the philosophical argumentation toward the naturalness of the theories involved, as we have seen that unnatural counterexamples are not difficult to find. As such, we'll focus on three relatively simple examples from Hamkins' paper that I think offer the most compelling philosophical defenses. The first is based on the idea of using some sort of calculation to find good axioms. The second is based on an alternative approach to enumerating our theories that takes care not to add dubious axioms. The third generalizes and addresses a problem with the first proposal by restricting our attention to the kinds of model that set theorists are generally more interested in. We shall be critical of each of these proposals and provide reasons why none of them is successful in establishing naturalness. However, I'd like to stress that despite the fact that Hamkins' paper is the apparent target of these criticisms, our real target sits in the background. There have been very few serious attempts to reject CHT, so Hamkins' paper should be recognized – at the least – for providing valuable clarification of just what is at stake. Our analysis of Hamkins' work will put us in a position to put CHT itself on trial in the final section of this paper.

2.2.1. Computing our axioms

Given that the essential objective of these counterexamples is arguably philosophical, it seems germane to begin with a kind of thought experiment. Suppose at some time in the not too distant future, most pure mathematicians have come to incorporate the use of computers into their mathematical practice. Even in the pursuit of most pure mathematics, many an hour is spent trudging through routine calculations that must be checked but yield little in the way of insight. The mathematicians of this future time have come some way to freeing themselves from these bonds. Now suppose that set theorists have come to consider a certain class of interesting mathematical problems; and they have figured out that for just about every problem in this class, there is a particular number $n \in \omega$ such that ZFC plus n inaccessible cardinals is, somehow, the optimal theory to address that problem. Supposing that the underlying class of problems is of independent mathematical interest, we would have reason to think that each of these extensions of ZFC is natural. Indeed, this move is arguably unnecessary since extensions of ZFC by inaccessible cardinals are almost universally regarded as natural theories. But suppose also that the calculation of the number of inaccessible cardinals appropriate to a particular problem is very tedious to calculate. Perhaps it involves a seemingly endless

[27] For another analysis and response to Hamkins' examples see (Grotenhuis, 2022). This dissertation takes up a Lakatosian approach that aims to use Hamkins' examples as a means to better isolate what a natural theory is. I am more pessimistic about the prospect of any satisfying analysis of natural theories, but this dissertation offers an intriguing point of view on this problem.

sequence of cases all based on a simple trick. The mathematicians of this fictional time knows what to do: they design a computer program that takes the statement of a problem from the relevant class and then after some time, returns the number of inaccessible cardinals required. This might result in a partial computable function from the naturals (as Gödel codes of statements of problems) to the naturals (as number of inaccessible cardinals to add to ZFC).[28] Moreover, the outputs of this partial function straightforwardly deliver natural extensions of ZFC. Let $\psi : \omega \rightharpoonup \omega$ denote such a partial computable function. Then for all $n \in \omega$

$$ZFC + \text{there are } \psi(n) \text{ inaccessible cardinals}$$

gives us a natural theory whenever $\psi(n)$ halts.

With our thought experiment complete, we now have an argument for the claim that we have a collection of natural theories. So far so good, but how to these theories play with CHT? While we have some comments to make in a moment, I think it is only fair to give Hamkins the mic dropping moment his intriguing result deserves.

Theorem 6. *(Hamkins, 2025) There is a partial computable function $\psi : \omega \rightharpoonup \omega$ such that the theories*

$$ZFC + \text{“there are } \psi(n) \text{ inaccessible cardinals.”}$$

for $n \in \omega$ have pairwise incomparable consistency strengths, assuming there are infinitely many inaccessible cardinals.[29]

Speaking loosely, this theorem appears to tell us that if we use the methodology suggested in the thought experiment above, then we could land in a situation were we have infinitely many natural theories, none of whose consistency strengths can be compared. As such, we have a much more serious challenge to CHT than we saw in Section 2.1. We seem to have both natural theories and nonlinearity.

I now want to push back against this putative counterexample from a couple of related angles. First, I'd like to consider how the theorem above is proved. Without getting too deep among the weeds, the essence of the proof is in finding the right partial computable function $\psi : \omega \rightharpoonup \omega$. The ψ required is very special: it is such that for any $m \neq n \in \omega$ there will be a model of ZFC where $\psi(m) < \psi(n)$ holds and another where $\psi(m) > \psi(n)$. Thus very informally, there are models of ZFC where one theory is stronger than the other and other models where the converse occurs. The rest of the argument is relatively straightforward. In order to obtain such a remarkable ψ, Hamkins makes an ingenious argument deploying the recursion theorem,[30] which states that for any total computable function $g : \omega \to \omega$ there is some $e \in \omega$ such that the partial computable function determined by e is the same as that determined by $g(e)$; or more formally, $\varphi_{g(e)} \simeq \varphi_e$. For the uninitiated, the recursion theorem is a close cousin of the notorious diagonal lemma used in proving Gödel's first incompleteness theorem. In both cases, we obtain a kind of fixed point through the use of a technical device simulating something like self-reference. The reader will recall that this was one of our superficial indicators of unnaturalness in a theory.

However, ψ has another remarkable feature that prompts further questions regarding its naturalness: the ψ used in the theorem never halts on any input in an ω-model! Indeed, this

[28] Recall that we merely assumed that *just about every* problem in the class had a number of inaccessible cardinals associated with it. Thus, the function could be properly partial rather than total.

[29] In fact, Hamkins proves a stronger results but this is more than enough for our purposes.

[30] For more details on the history of this kind of argument and some related theorems see (Hamkins, 2025).

fact is crucial for the proof of its special properties. Doesn't it seem a little odd to call a theory natural when – modestly assuming the actual natural numbers are well-founded – that theory is not properly defined? Moreover, if we return to the thought experiment that began this section, why would we select such a function for such a job? Hamkins is, of course, aware of this issue and offers an interesting response that I'll now try to motivate.[31] Suppose that in our thought experiment, our problems correspond to slightly different extensions of ZFC. Rather than saying that "there are $\psi(n)$ inaccessible cardinals," we might instead say "there are *at most* $\psi(n)$ inaccessible cardinals." So the natural theory returned puts a kind of bound on the number of inaccessible cardinals appropriate to the problem. Now when we feed ψ the code n of some problem, and the calculation of $\psi(n)$ does not terminate, an obvious interpretation suggests itself. The calculation doesn't terminate since there is no finite bound on the number of inaccessible cardinals appropriate to the problem at hand. Thus, when $\psi(n)$ doesn't terminate we could interpret this as determining the theory of ZFC with infinitely many inaccessible cardinals appended to it. I think this response is coherent and fits neatly with the philosophical motivations of these theories. So there is some reason to think that the naturalness of the initial thought experiment is preserved. But I also think this response makes such theories seem unnatural on other grounds. Suppose that given the considerations above, we countenance the use of partial computable functions in determining theories and further, we assume that theories are still determined even when that partial computable function does not halt. In the example above, we can prove that ψ never halts, so we know what to do. But as a general principle, this seems very strange. In general, this will lead us into positions where we are describing theories whose axioms cannot – even in principle – be determined by finitary means. Traditionally, foundational theories are prized for the simplicity of their axiomatization. While ZFC isn't finitely axiomatizable, it's arguably the next best thing. There's a finite set of very simple instructions that anyone can follow to determine whether *or not* some formula is an axiom of the theory. Without at least this, it's difficult to understand how we could – as finite beings – make use of such a theory. But this is what Hamkins is offering us here. The crucial point is that in a computable axiomatization, we use a total computable function, while Hamkins admits partial computable functions and assigns them values even when they do no halt. The only way for us to competently use such a theory would be through an impossible solution to the halting problem. I think this gives us a more substantive reason to doubt the naturalness of such theories.

My final quibble with this example is more subtle, but I also think more pressing. The issue this time is metamathematical and concerns how the theorem above should be stated in order to be in accord with our informal interpretation of it. The thinking here reminds me of the way in which a great painting often demands that its viewer think about where the artist was standing when they produced the image. Similarly in this case, we need to think about where we are standing when stating Theorem 6. In particular, I am concerned with where the computer, which figures out how many inaccessible cardinals to add to our theory, is being operated. Is it here with us, or in some other abstract context? My contention is that this computer is not being operated here, but rather in some abstract, nonstandard model of set theory that is quite confused about which computations terminate and which do not. The easiest way to see this is that in our world, the crucial function ψ never halts on any input that it is given. Thus, if the

[31] I do things a little differently to Hamkins in order to keep with the flow of this discussion, but I hope to have captured the spirit of his response.

calculation of $\psi(n)$ for $n \in \omega$ were undertaken here, then we'd just learn that for all $m < n \in \omega$

$$ZFC + \text{“there are } \leq \psi(m) \text{inacessibles”} \equiv_{Con} ZFC + \text{“there are } \leq \psi(n) \text{inacessibles”}$$

since neither $\psi(m)$ nor $\psi(n)$ are defined and so they both say that there are infinitely many inaccessibles.[32] I think this interpretation is in good accord with the thought experiment from the beginning of this section. If we want to check how many inaccessibles to add to ZFC in order to align with some problem, then it makes good sense to use a computer if the required calculation is long and tedious. But we want the computer to do essentially what we would have done, just without our having to do it; we don't want it to operate in some nonstandard environment and return a value when no such value is due. The theory we are interested in, is the theory that the calculation delivers in this world. But the crucial trick of Hamkins example relies on doing the exact opposite of this. We take the computer program and we run it in models where it does halt, but after a nonstandard number of steps. These are the only models where we get $\psi(m) < \psi(n)$ and $\psi(m) > \psi(n)$. But in such models

$$ZFC + \text{“there are } \leq \psi(m) \text{inacessibles”}$$

has a completely different meaning that it does here in our world. In those worlds, it says that there at most $\psi(m)$ inaccessible cardinals, while here it says there are infinitely many. More succinctly, we might put the problem as follows: we set out to compare two theories that interested us and ended up comparing two completely different theories instead.[33]

Before we move to the next example, I want to take a more strategic perspective on what we've seen above. While I've offered some criticisms of Hamkins' argument toward the naturalness of the theories he proposes, I don't want want to pretend that my criticisms are decisive refutations. This is, in part, because I have a congenial dialectical position in this debate. Even though it was quite vaguely defined, naturalness is a high bar for a theory to meet. As such, to push back against Hamkins' example it would probably suffice to just show that the water is muddy and that plausible controversy lurks at every turn. This would be enough to tarnish the credentials of a putative natural theory. While I think we have done more than this above, I think it is important to note which way the deck is stacked in this game. Indeed, I think this should have some effect on how we evaluate the significance of the Consistency Hierarchy Thesis.

2.2.2. Being very careful

For our next example, we again begin with a thought experiment to motivate the claim that a proposed extension of ZFC is natural. As we noted at the beginning of this paper, Gödel's

[32] Here we are using the fix for cases where ψ doesn't terminate. If we didn't use that, we wouldn't have any theories at all.

[33] It's important to note that this point is quite dependent on the thought experiment suggested above. As such, there could well be another way of motivating something like the theory above that is not exposed to this style of objection. To highlight the dependence, we note that ZFC itself is a computably axiomatizable theory that is interpreted quite differently in other worlds. For example, if we suppose that $ZFC \nvdash Con(ZFC)$, then by completeness we may fix a model $\mathcal{M}$ satisfying $ZFC \cup \{\neg Con(ZFC)\}$. However, the sentence $\neg Con(ZFC)$ is interpreted quite differently in $\mathcal{M}$ than in our background universe. This is because the statement $\neg Con(ZFC)$ requires the use of a formula to represent the codes of sentences from ZFC. Then since $\mathcal{M}$ satisfies $\neg Con(ZFC)$, $\mathcal{M}$ will think some largest natural number x such that the axioms of ZFC whose codes are below x are consistent. But the Reflection Theorem implies that every truly finite subset of ZFC is consistent in $\mathcal{M}$. This means that x must be a nonstandard element of $\omega^{\mathcal{M}}$. But then we we apply the Reflection Theorem inside $\mathcal{M}$ to the axioms whose codes are below x, we get a model $\mathcal{N}$ satisfying every genuine axiom of ZFC. So there is a sense in which $\mathcal{M}$ thinks ZFC is consistent after all. See the discussion on page 146 of (Kunen, 2006) for a more patient rendering of this argument. The upshot though is that the way in which theories dependent on computation are interpreted can be very dependent on context. Intuition certainly seems to struggle. This may give us a different reason to doubt that we are doing something natural and indeed later, in Section 3.2., this will give us some reason to reevaluate how we should interpret such theories.

incompleteness theorems can be understood as leaving open a chaotic realm beyond ZFC. While we saw that CHT goes some way toward taming this menagerie, Gödel's theorems also prompt another worry. To see this, first recall that ZFC is sometimes thought to delimit what we can assume, without remark, in a mathematical publication. If you can prove φ in ZFC, then φ is just a theorem in your paper; but if you also need a measurable cardinal, then your theorem is: φ, if there is a measurable cardinal. Roughly speaking, we might say that ZFC sets out the limits of consensus mathematical knowledge. But then by this standard, Gödel's second theorem would tell us that, for all we know, ZFC is inconsistent since this cannot be proved in ZFC, and we are all just wasting our time. Of course, there have been many attempts to argue on different grounds for the consistency of ZFC. Perhaps the best of them relies on the fact that – as of now – no proof of inconsistency has been found from the axioms of ZFC in over a century of use. But for those of a more anxious temperament, such empirical fodder could come as cold comfort. This is where Hamkins' next proposal enters our story.

> ... let the *cautious enumeration of* ZFC be the enumeration of ZFC that continues as long as we have not yet found a proof in what we have enumerated so far that ZFC is inconsistent. I denote the resulting theory by ZFC^o. In order to halt the enumeration we don't require an explicit contradiction in ZFC, but rather only a proof that there is such a contradiction (Hamkins, 2025).

Thus, rather than naively admitting all of ZFC into our mathematical knowledge base, we also perform an extra safety check with the goal of obtaining greater confidence in our theory. At face value, this extra move seems prudent and sensible. As such, if we can formulate a theory that achieves this effect it seems reasonable to call it natural. I now want to provide a more detailed exposition of how to understand the cautious enumeration. We will then be able to generalize this cautious position and obtain a collection of theories that form an infinite descending chain in the consistency hierarchy and thus, a further challenge to CHT.

We start with some general remarks about notation and enumeration of theories.[34] We shall dive a little deeper into the details than in the previous section. This is the most technical section of the paper and the reader may be best served by glossing it on a first reading. Following Lindström (2003), let us take it that ZFC denotes its axioms rather than, the more traditional closure of the axioms in first order logic. Suppose then that $\ulcorner\cdot\urcorner : \mathcal{L}_\in \to \omega$ is a Gödel coding; i.e., a computable injection from the formulae of set theory into the natural numbers. Now we know that there is a Σ^0_1 formula[35] $\tau(x)$ that enumerates the axioms of ZFC in such a way that for all $\varphi \in \mathcal{L}_\in$

$$\varphi \in ZFC \;\Leftrightarrow\; V_\omega \models \tau^{\ulcorner}\varphi^{\urcorner}$$

and further

$$\varphi \in ZFC \;\Rightarrow\; PA \vdash \tau^{\ulcorner}\varphi^{\urcorner}$$

and

$$\varphi \notin ZFC \;\Rightarrow\; PA \vdash \neg\tau^{\ulcorner}\varphi^{\urcorner}.$$

We have this since ZFC is computably axiomatizable and $\tau(x)$ witnesses this. Moreover, we shall suppose τ works by simply identifying the codes of the axioms which are not schema and

[34] In general, we aim to follow a slightly modernized version of the notation and terminology of (Lindström, 2003).

[35] Since we are working in set theory, we shall think of a Σ^0_1 formula a Σ_1 formula of the Lévy hierarchy that is then relativized to V_ω. See Kunen (2009) for a thorough treatment of this approach.

for Separation and Replacement it employs some very simple form of pattern recognition. We shall also say that $\tau(x)$ *binumerates* ZFC and we shall fix such a τ for the remainder of this section. For an arbitrary Σ_1^0 formula $\sigma(x)$, we shall let $B_\sigma{}^\ulcorner\varphi^\urcorner$ be a Σ_1^0 statement saying that there is a natural number coding a proof of φ using codes of formulae ψ such that $\sigma^\ulcorner\psi^\urcorner$ holds. We then write $Con(\sigma)$ to mean $\neg B_\sigma{}^\ulcorner\emptyset \neq \emptyset^\urcorner$; i.e, σ defines a set of axioms that are consistent. Thus, where we've written $Con(ZFC)$, we may now write $Con(\tau)$.[36] Let $Bew_\sigma(^\ulcorner\varphi^\urcorner, d)$ mean that d is the code of a derivation of φ from assumptions whose codes satisfy $\sigma(x)$. These technical considerations are important in this case since they are essential to our definition of an enumeration. For our purposes, $\tau(x)$ determines an enumeration of ZFC in the form of an ordering $\prec_\tau$ such that for all formula φ, ψ of $\mathcal{L}_\in$

$$\varphi \preceq_\tau \psi \;\Leftrightarrow\; \tau^\ulcorner\varphi^\urcorner \wedge \tau^\ulcorner\psi^\urcorner \wedge {}^\ulcorner\varphi^\urcorner \leq {}^\ulcorner\psi^\urcorner.$$

Thus informally, φ precedes ψ if they are both axioms of ZFC and the code of φ precedes that of ψ. Finally, we include a technical notation that will be helpful in articulating Hamkins' cautious enumeration. For arbitrary $\sigma(y)$ from $\mathcal{L}_\in$, let $(\sigma|x)(y)$ be the formula: $\tau(y) \wedge y \leq x$. Thus, $(\sigma|x)(y)$ defines the initial segment determined by $\preceq_\sigma$ below x.

With these remarks out of the way we are ready to provide a formalization of Hamkins' proposal by letting the cautious enumeration be determined by the formula $\tau^o(x)$, which says

$$\tau(x) \wedge \forall d < x\, \neg Bew_{\tau|x}(^\ulcorner\neg Con(ZFC)^\urcorner, d).$$

Less formally, $\tau^{o\ulcorner}\varphi^\urcorner$ holds just in case φ is an axiom of ZFC and "we have not yet found a proof" in the sense of there being some $d < x$ coding a proof of $\neg Con(ZFC)$ from axioms among those "we have enumerated so far" in the sense that they are from $\tau|x$. I think this interpretation gives a reasonably faithful reading of what Hamkins means when he speaks of "what we have enumerated so far" and what it means to "have not yet found a proof." Morally speaking, if we ever considered adding to our stock an axiom from which we can prove the inconsistency of ZFC, then we have reason to stop the enumeration. On this basis, we might argue for its naturalness.

Our next goal is to generalize this rendering of caution and obtain a failure of well-foundedness. But before we do this, we highlight the key result for this example. We first recall the well-known fact that ZFC is essentially reflexive.

Fact 7. *(ZFC) If T is a theory extending ZFC in $\mathcal{L}_\in$ then for all finite subsets Δ of T, $ZFC \vdash Con(\Delta)$.*

This is a straightforward consequence of the reflection theorem and it allows us to establish the following.

Theorem 8. *(Hamkins, 2025) $Con(\tau^o) <_{Con} Con(ZFC)$, supposing $Con(ZFC + Con(ZFC))$.*

Although this proof occurs in (Hamkins, 2025), we are going to make a few simple generalizations of it below so it will be valuable to have a suitable version of the proof available for inspection.

[36] I'm going to continue to write $Con(ZFC)$ in deference to standard conventions. However in contrast to Hamkins, I will use the $Con(\sigma)$ notation for the more exotic enumerations introduced by Hamkins and discussed below.

Proof. Since $\tau^o \subseteq ZFC$, it is obvious that $Con(\tau^o) \leq_{Con} Con(ZFC)$, so it suffices to show there is a model $\mathcal{M}$ satisfying $Con(\tau^0)$ but $\neg Con(ZFC)$. By our assumption and Gödel's second theorem, we may fix a model $\mathcal{N}$ satisfying

$$ZFC + Con(ZFC) + \neg Con(ZFC + Con(ZFC)).$$

Then in $\mathcal{N}$, we have $ZFC \vdash \neg Con(ZFC)$. Thus, we see that τ^o is a finite subset of ZFC according to $\mathcal{N}$. Now using Fact 7 inside $\mathcal{N}$, we see that $\mathcal{N}$ also thinks that $ZFC \vdash Con(\tau^o)$. Then since $Con(ZFC) \rightarrow Con(ZFC + \neg Con(ZFC))$ holds in $\mathcal{N}$, we see that $\mathcal{N}$ also thinks $Con(ZFC + \neg Con(ZFC))$ and so we may fix a model $\mathcal{M}$ in $\mathcal{N}$ satisfying $ZFC + \neg Con(ZFC)$. But since $\mathcal{N}$ also thinks $ZFC \vdash Con(\tau^o)$, we see that $\mathcal{M}$ satisfies $Con(\tau^o)$ and $\neg Con(ZFC)$ as required. □

Thus and perhaps surprisingly, we see that the cautious enumeration ends up having a consistency strength strictly below ZFC as it is ordinarily enumerated. With this, the infinite descent begins. For the next step, we adopt an even more cautious position and check that – so far – not only have we failed to find a proof that ZFC is inconsistent, we have also not found a proof that $ZFC + Con(ZFC)$ is inconsistent. More formally, we let $\tau^{oo}(x)$ be

$$\tau(x) \wedge \forall d_0 < x \, \neg Bew_{\tau|x}(\ulcorner \neg Con(ZFC) \urcorner, d_0) \wedge \forall d_1 < x \, \neg Bew_{\tau|x}(\ulcorner \neg Con(ZFC + Con(ZFC)) \urcorner, d_1).$$

Call this the *doubly cautious enumeration*. Essentially the same proof then gives us the following.

Theorem. *(Hamkins, 2025)* $Con(\tau^{oo}) <_{Con} Con(\tau^o)$*, supposing*

$$Con(ZFC + Con(ZFC) + Con(ZFC + Con(ZFC))).$$

And from here it is not difficult to see that the we can generalize this to form triply cautious enumerations, quadruply cautious enumerations and so on.[37] Then the proof of Theorem 8 can be easily adapted to obtain the following.

Theorem. *(Hamkins, 2025) For all* $n \in \omega$, $Con(\tau^{o(n+1)}) <_{Con} Con(\tau^{o(n)})$*, supposing for all* $n \in \omega$

$$Con(ZFC^{o(n)}).$$

Taking some stock, we started out by just wanting to be a more careful in the face of the ever-present risk of inconsistency associated with ZFC and its extensions. This led us to adopt a more conservative approach to the enumeration of ZFC with the goal of mitigating at least some of that risk. On the basis of the sensibleness of this worry, we have argued that the resulting theories are natural. But as we see above, they also form an infinitely descending chain of consistency strengths and thus a challenge to the CHT.

Such is Hamkins' second proposal. The mathematics is undeniable, so as in the previous section, we will now attempt to push back on the more philosophical claim that these theories are natural. We shall consider two objections. Our first objection begins with what might be

[37] More formally, Let $ZFC^{o(0)}$ be $\tau(x)$; i.e., ZFC. Let $ZFC^{o(n+1)}$ be $ZFC^{o(n)} + Con(ZFC^{o(n)})$. Let $\tau^{o(0)}(x)$ be $\tau^o(x)$ and let $\tau^{o(n+1)}(x)$ be

$$\tau^{o(n)}(x) \wedge \forall d_n < x \, \neg Bew_{\tau|x}(\ulcorner \neg Con(ZFC^{o(n+1)}) \urcorner, d_n).$$

better construed as a criticism of my formal interpretation of the cautious enumeration. While Hamkins' is informal, I characterized this enumeration using the formula $\tau^o(x)$ which says

$$\tau(x) \wedge \forall d < x \, \neg Bew_{\tau|x}(\ulcorner \neg Con(ZFC) \urcorner, d).$$

While we arguably have more precision, it also invites a couple of pointed questions:

(i) Why should we just consider proof codes that precede x according to the standard ordering of the naturals in this particular coding?
(ii) Why should we consider initial segments of τ determined by the standard ordering of the naturals?

I think there are easy ways to modify τ^o that address these questions and the ensuing discussion will serve to illustrate how our thought experiment above should be understood in more detail. With regard to (1), it does seem a little odd to just use proof codes that precede x according to $<$. Why use x? Why use $<$? With regard to using x, we use it simply because that's the only variable available. But perhaps we want to check more than just those proofs below x. To do this, we could simply take any computable, order-preserving map $f : \omega \to \omega$ and use $f(x)$ rather than x as the bound. With regard to $<$, this is the same problem underlying (2). Why should our enumeration be hostage to the standard ordering of the naturals as mediated by some Gödel coding? This is also easy to address by fixing a computable permutation $g : \omega \to \omega$ and using the order $\prec_g$ instead, where for all $m, n \in \omega$

$$m \prec_g n \;\Leftrightarrow\; g(n) < g(m).$$

If we we make these modifications let $\tau^o(x)$ say

$$\tau(x) \wedge \forall d \prec_g f(x) \, \neg Bew_{\tau|_g x}(\ulcorner \neg Con(ZFC) \urcorner, d)$$

where $(\tau|_g x)(y)$ says $\tau(y) \wedge y \prec_g x$. This again results in a computable axiomatization of the theory in question. None of this affects the results above. Moreover, this arguably helps us show how the cautious enumeration fits our motivating story a little more tightly. We might think of $\prec_g$ as giving the enumeration of the axioms as we come to use them. So if we come to use some instance of Replacement after an instance of Separation, then this could be reflected in the $\prec_g$ ordering.[38] Similarly, we might think of f as telling us how many proofs we worked out when we added the new axiom.[39] Arguably this describes something like the process of set theorists working today. Thus, while I think it is fair to say that my original formalization of τ^o was somewhat crude, it is not difficult to patch things to be in better accord with our underlying motivations.

But I think this discussion prompts a reasonable question that leads to an objection due to John Steel. The underlying idea behind the cautious enumeration is a prudential attitude in response to the epistemic hurdles erected by Gödel's second theorem. So what happens if the cautious enumeration gets stuck? Suppose we add a new axiom and do a few proofs and establish on the basis of what we have so far that $\neg Con(ZFC)$. What should we do? This is not an outright proof of inconsistency, so it doesn't quite tell us that what we have is trivial. Nonetheless it should give us reason for pause. And this is exactly what the cautious

[38] This assumes that there is a computable permutation giving the order of discovery, but that seems like a relatively plausible assumption.
[39] Strictly, I think it would be more faithful to introduce a further computable permutation $h : \omega \to \omega$ to deliver another ordering $\prec_h$ that records the order of discover of proofs.

enumeration tells us to do. But what next? Since we discovered the moment when the rot set in, perhaps we should retreat to what we had just before that occurred. If we did this, then we'd still in accord with τ^o so this seems to line up with our motivating story. But, and here is where the objection comes in, this doesn't seem to line up with what set theorists would actually do in such a situation. While we can agree that the proof of $\neg Con(ZFC)$ is a good reason to pause, the sensible response would be to undertake a much deeper kind of retreat. Steel puts it as follows:

> If we discovered a proof of $\neg Con(ZFC)$, we wouldn't retreat to "the amount of ZFC enumerated so far." We'd say the ideas were wrong, and drop back based on some analysis of how the ideas went wrong.[40]

In other words, the occurrence of such an event would shake us back to the foundations. Such a result would force us to see that there was something deeply wrong in our understanding of set theory and this misunderstanding would rightly prompt questions about our entire axiomatization. Given this, τ^o and its cousins might seen less natural than at first blush.

For our second objection, we focus on the role of finitude in the results above and reflect on how this injects a kind of strangeness into the τ^o enumeration. First recall the crucial role of ZFC's essential reflexivity (Fact 7) in the proof of Theorem 8. It allowed us – upon recognizing that τ^o had a finite extension – to fix a model satisfying τ^o. We also observed above that essential reflexivity was a simple corollary of the Reflection Theorem in set theory. We shall demonstrate below that there is a sense in which the theory enumerated by τ^o doesn't satisfy the Reflection Theorem, which seems like particularly unnatural feature in a theory given the technical and philosophical importance of that theorem.

The key point is that assuming ZFC is consistent, we cannot prove from the axioms enumerated by τ^o whether τ_0 determines a finite axiomatization or not. As such, we aren't in a position to know whether reflection is available or not. More formally, we have:

Proposition 9. $\neg B_{\tau^o}\ulcorner\neg\exists x\forall y((\tau^o|x)(y) \leftrightarrow \tau^o(y))\urcorner$, *if* $Con(ZFC)$.

Less formally, this says that we cannot prove from axioms satisfying τ^o that: there is no finite initial segment of τ^o indexed by some x that is the entirety of τ^o, assuming ZFC is consistent.

Proof. Since we have $Con(ZFC)$, we may use Gödel's second theorem to fix a model $\mathcal{M}$ satisfying $ZFC + \neg Con(ZFC)$. Now as in the proof of Theorem 8, $\mathcal{M}$ will think that there are only finitely many code numbers satisfying τ^o; i.e., $\exists x\forall y((\tau^o|x)(y) \leftrightarrow \tau^o(y))$. Thus since $\mathcal{M}$ satisfies the axioms that satisfy τ^o, we see by soundness that we cannot prove $\neg\exists x\forall y((\tau^o|x)(y) \leftrightarrow \tau^o(y))$ from axioms satisfying τ^o. □

This in itself is an unusual feature of the τ^o enumeration. I think this tells us that any reasonable enumeration of ZFC should not be like this. Indeed, it is easily proven using the Reflection Theorem that ZFC cannot be generated from any finite subset of itself, provided that ZFC is consistent; i.e., i.e., ZFC is not finitely axiomatizable. Indeed one can prove this

[40] Quoted with permission from an email between John Steel and me.

in a theory of arithmetic like PA.[41] Or more formally, recalling our very simple enumeration τ, described at the beginning of Section 2.2.2., we have:

Proposition 10. *Suppose that ZFC is consistent. Then*

$$B_\tau{}^\ulcorner\neg\exists x\forall y((\tau|x)(y) \leftrightarrow \tau(y))^\urcorner.$$

Thus, if we use an enumeration like τ, then even a very weak theory can prove that τ determines an infinite theory. Moreover, it is easy to see that this result perseveres to stronger theories using the obvious modifications. The essential ingredient of this proof is, of course, the Reflection Theorem. As such, we now have some good reason to doubt that τ^o determines a theory with the Reflection theorem. We close this section by spelling this out in some detail.

We start with a way of defining theories that satisfy the Reflection Theorem. It is traditional to state the Reflection Theorem as a schema and then prove it as a kind of meta-theorem. But as our concerns here are particularly metamathematical, we add the fussy detail of a metatheory (PA) from which the theorem can proved.

Definition 11. Let us say that a formula $\sigma(x)$ of $\mathcal{L}_\in$ *enumerates a reflective theory* if PA proves that whenever $\sigma^\ulcorner\varphi_0{}^\urcorner \wedge \ldots \wedge \sigma^\ulcorner\varphi_n{}^\urcorner$ then[42]

$$B_\sigma{}^\ulcorner\forall\alpha\exists\beta > \alpha \bigwedge_{i<n} \varphi_i^{V_\beta}{}^\urcorner.$$

Thus, we have a reflective theory when PA can prove that any finite set of sentences satisfying $\sigma(x)$ are themselves satisfied in a rank initial segment of the universe. This is, of course, a well-known and weak version of the Reflection Theorem. A standard argument then shows that such theories cannot be finitely axiomatized.

Lemma 12. *If σ binumerates a consistent reflective theory, then PA proves that there is no x such that*[43]

$$\forall y(\sigma(y) \leftrightarrow (\sigma|x)(y)).$$

We can then put all this together to see that τ^o doesn't satisfy the Reflection Theorem.

Theorem 13. *τ^o does not enumerate a reflective theory, if ZFC is consistent.*

Proof. Work in PA and suppose toward a contradiction that τ^o does enumerate a reflective theory. Then Lemma 12 and Proposition 9 with $Con(ZFC)$ respectively imply that there is and there is not some x such that $\forall y(\sigma(y) \leftrightarrow (\sigma|x)(y))$, which is impossible. □

Summing up, we see that while Gödel's second incompleteness theorems reasonably prompt concerns about the consistency of strong theories like those extending ZFC, the effort to ameliorate these worries by modifying the way we enumerate axioms leads into a couple of arguments against the naturalness of such theories. First, we argued via Steel that the cautious

[41] In fact primitive recursive arithmetic would suffice. See Corollary IV.7.7 in (Kunen, 2006) for a proof and further discussion of the metamathematical setting. For the purposes of uniformity, we shall let PA be the set theory formed by: removing the axiom of infinity; adding its negation; and then replacing Foundation with Set Induction. The resultant theory is well-known to be bi-interpretable with the standard version of PA so no harm has been done (Kaye and Wong, 2007).

[42] This is essentially says that $\sigma(x)$ determines a theory satisfying a fussy restatement of Corollary IV.7.6 of (Kunen, 2006).

[43] A proof of this essentially follows that of Corollary IV.7.7 in (Kunen, 2006).

enumeration doesn't fit mathematical practice as well as it might seem as first. Second, we see that properties we have come to take for granted with regard to ZFC are no longer guaranteed in the context of unorthodox enumerations. Arguably the real lesson here is that we should be more careful in stating what we really want from an enumeration of a theory like ZFC. Computability is certainly a necessary condition, but each of the enumerations above are computable axiomatizations of ZFC and yet they behave very oddly.

Before we move to our final example, I want to stress again that I doubt that I've definitively argued that the enumerations discussed in this section are unnatural. As in the previous section, I wouldn't be surprised if one could push back. However, with every passing epicycle in such a debate, I think the case for unnaturalness looks stronger for the simple reason that a natural theory should be obviously so. This again is a testament to my dialectical position and should cast no shade on Hamkins' innovative examples.

2.2.3. Sticking to the good models

This final example is a variation of the one from Section 2.2.1., but I think it's important to discuss since it scratches at an itch that many set theorists will be feeling right now. While relative consistency proofs like Gödel's first incompleteness theorem are of great interest, relative consistency proofs in set theory tend to demonstrate a much tighter relationship between the theories in question. To see this, let's consider a well-known forcing example. Suppose we want to show that $ZFC + \neg CH \leq_{Con} ZFC$. To do this we might start with a countable transitive model M satisfying ZFC and then show that there is an M-generic set G such that the extension $M[G]$ of M safisfies ZFC and $\neg CH$. Thus, we have shown that if there is a transitive model of ZFC, then there is a transitive model of $ZFC + \neg CH$.[44] Let's introduce a little notation for this by writing $Con_\beta(T)$ to mean that there is a transitive model of T.[45] And let us write $T \models_\beta \varphi$ to mean that every transitive model of T also satisfies φ. Then we have shown that

$$Con_\beta(ZFC) \rightarrow Con_\beta(ZFC + \neg CH).$$

Analogously, we may also show that

$$Con_\beta(ZFC) \rightarrow Con_\beta(ZFC + CH)$$

using Gödel's inner model argument.[46] Now since we have restricted our attention to countable transitive models, we have not quite demonstrated that $ZFC + \neg CH \leq_{Con} ZFC$. There are a number of standard methods of patching up the proof to achieve this,[47] but this is the conceptual core of the argument. Moreover, this way of looking at things reveals a much closer connection between the theories in question. For example, the models of each theory have the same ordinal spine. By contrast, we cannot prove that

$$Con_\beta(ZFC) \rightarrow Con_\beta(ZFC + \neg Con(ZFC))$$

even though $ZFC + \neg Con(ZFC) \leq_{Con} ZFC$. This is because if there is a countable transitive model of ZFC, then no model of $ZFC + \neg Con(ZFC)$ can be transitive since transitive models have the true set of natural numbers and so they must see that $Con(ZFC)$ is true. So we start

[44] Note that if there's a transitive model, then there is a countable transitive model too.
[45] The idea here is that T has a β-model, where a β-model of $\mathcal{L}_\in$ is well-founded. Then since we are working in ZFC as a background theory, we may assume that such a model is transitive without loss of generality.
[46] Of course, there are no issues regarding countable models in this case.
[47] See Section VII.9 of (Kunen, 2006) for an excellent overview.

with a good, well-founded model of ZFC and then obtain an ill-founded model whose ordinals have no meaningful relationship with the ordinals of the model we started with.

Thus, forcing and inner models seem to tell us much more about how the theories they discuss are related than a mere relative consistency argument. Moreover, as we saw in Section 2.2.1., these considerations give us reason to worry about the naturalness of theories extending ZFC that are not within the reach of forcing or inner model theory. At the least, we would hope that the existence of a transitive model of one natural theory should imply the existence of a transitive model of another. With this in mind, let us introduce a hopefully better strength relation between theories.

Definition 14. For theories T and S extending ZFC in the language of set theory, let us say that T is *transitively consistent relative to* S, abbreviated $T \leq_{trans} S$ if

$$ZFC \vdash Con_{\beta}(S) \to Con_{\beta}(T).$$

Given that we have just seen that $ZFC + \neg Con(ZFC) \not\leq_{trans} ZFC$, we have some reason to hope that the kinds of counterexamples that we have seen above will be eradicated and that CHT will be vindicated in this context. But intriguingly, Hamkins shows that this is not the case. Using essentially the same proof strategy as he employed for Theorem 6, he is able to show the following.[48]

Theorem 15. *(Hamkins, 2025) There is a partial computable function $\psi : \omega \rightharpoonup \omega$ such that the theories*

$$ZFC + \text{“}\textit{there are } \psi(n) \textit{ inaccessible cardinals.}\text{”}$$

for $n \in \omega$ are pairwise $\leq_{trans}$ incomparable, assuming there are infinitely many inaccessible cardinals.

Thus, even if we restrict our attention to transitive models, we have the mathematical bones of a counterexample to CHT. Of course, many of the worries raised in Section 2.2.1. are directly transferable to the current context: we are axiomatizing a theory using a partial computable function, which makes figuring out the axioms very difficult; and we seem overly concerned with the behavior of a partial computable function in nonstandard contexts where infinite natural numbers reside. As such, I still think we have good reason to be reserved about calling such a theory natural, but the generalization to transitive models is certainly remarkable and is a significant milestone for the discussion that is to come.

3. Explaining the difficulty

Thus far, we have explored the mathematical and philosophical significance of the Consistency Hierarchy Thesis and considered a number of putative counterexamples with regard to which we have expressed some reservations. In particular, we have raised doubts about the naturalness of the theories offered in these counterexamples. In this final section, I want to consider things from a different perspective. Rather than thinking about the actual examples that support the thesis and how some specific examples fail to defeat it, I want to consider the endgame of the problem itself. As such, I have two main goals. First, I want to think more seriously about what an uncontroversially successful counterexample to CHT would look like

[48] In fact, one might even say that the proof is a little easier.

and the likelihood of such a scenario taking place. Second, I want to pull together a few of the recurring threads of this paper and offer a kind of formal explanation as to why refuting CHT seems so improbable. My hope is to demonstrate that CHT is a somewhat idiosyncratic problem and that its peculiarities should be taken into account when we assess its significance.

3.1. The ideal counterexample scenario

In Section 2, we considered putative counterexamples to CHT that aimed to show that the consistency hierarchy for natural theories is not well-ordered by offering examples where comparability and well-foundedness fail, but we also saw some reasons to doubt that the examples really delivered natural theories. In this section, rather than attempt to provide a specific example, I want to take a step back and paint a picture of what I think the ideal counterexample to CHT would look like. My goal is to consider a scenario that would have maximal impact on the set theory community and also be accepted by them. I think there is value in exploring this idea as it makes clearer just how thorny this problem is.

We begin our story with young algebraic topologist, Gurt Ködel. Ködel has been working on a collection of seemingly natural problems in his field but has been unable to find any way to solve them. These problems seem like sensible generalizations of problems already considered by his mathematical community and yet they seem to lie beyond the reach of existing techniques. Naturally enough, he wonders whether these problems are perhaps independent of ZFC, but again, he is unable to get very far with this question. Nonetheless, Ködel does develop a new axiom (or perhaps collection thereof) which seems to add exactly the right ingredient to address the problems he has raised. Moreover, he discovers that this new axiom is able to address other problems, some of which were already in the literature and some of which are quite new. Ködel shares these results with his colleagues and goes on to gain a degree of fame by having opened the door to a hitherto unexplored region of mathematics. Other mathematicians go on to solve further problems with this axiom and this new field continues to flourish. One might imagine this taking place over a period of years or even decades.

This describes a prototypical scenario for the emergence of a natural theory. A new axiom is introduced for a clear mathematical purpose and even better, the development of the theory around this axiom garners wide interest in the mathematical community. Note that in contrast to the examples above, the story here is organic rather than contrived. Life is often easier in thought experiments.

So far so good. As the notoriety of this new field grows the set theory community gets wind of its success and ask the question: where does this new theory fit in the large cardinal hierarchy? What is its consistency strength? It is well-known that Ködel and his successors were unable to solve this problem and so it remained open. This raises further difficult philosophical questions. Perhaps the most important of these is: why is this a problem and who is it a problem for? On one hand, the set theory community has a pretty standard line on these matters: if you cannot prove that a theory is consistent relative to a large cardinal axiom, then we have no reason to think that theory is consistent. Perhaps the classic case of this is Quine's theory NF. While the jury might be out as to whether this really is a natural theory, there is not doubt that it currently lacks a consistency proof that has been agreed to be correct by the mathematical community.[49] As such, it is customary to regard the question of NF as open. However, if one were able

[49] Interestingly, during the time this paper was written NF now has a computer-verified consistency proof (Holmes and Wilshaw, 2024). This does not appear to be very well-known in the set-theory community yet, however, it seems reasonable to expect that there will now be consensus that the consistency

to show that NF was consistent relative to a large cardinal (or more likely much less) then we would regard the question of NF's consistency as having been resolved. It is important, however, to consider the significance of CHT in the set theorist's standard line. The success of CHT with regard to what we currently know is what gives us reason to think that consistency relative to a large cardinal axiom is – at least – a necessary condition for thinking that a theory is consistent. If CHT were to face serious doubts, then a failure to determine consistency relative to a large cardinal axiom would carry much less weight. But this is precisely the kind of pressure the natural theory imagined above would place on CHT. Our thought experiment is exploring a scenario where a theory governing an active area of mathematics cannot be understood as fitting into the large cardinal hierarchy.

Continuing our story, the consistency of Ködel's theory becomes one of the dominant open questions of set theory. Partial questions are developed and answered using forcing and inner model theory, but the main question remains unanswered. Much imaginative but failed work remains in notebooks and on blackboards never seeing the light of publication. The question becomes an albatross for CHT as work on Ködel's theory continues to flourish. And as is not atypical in mathematics, the question remains open for decades.

At this juncture in our imagined scenario, CHT looks much weaker than it does in our world. It hasn't been refuted, but the idea that the large cardinal hierarchy sets the milestones on the long road of theoretical strength is certainly tarnished. I think it's also worth noting that the kind of pressure placed on CHT in this example, is quite different to the examples considered in Section 2. In those cases, we considered examples of theories that refuted well-orderedness by breaking comparability and well-foundedness. Here, we are seeing a problem with a facet of well-orderedness that is arguably more significant for the philosophical and foundational consequences of CHT: that the consistency hierarchy might not be directed. Or more specifically, Ködel's theory challenges the idea that for any natural theory T there is a large cardinal axiom LC_T such that $T \leq_{Con} LC_T$. In other words, this would challenge the idea that the large cardinal axioms are cofinal in the hierarchy of consistency strengths. While breaking comparability or well-foundedness with natural theories, would certainly refute CHT as we have defined it above, one might easily design a fallback version of CHT that gives up on these features. While this would be disappointing, much of the philosophical and foundational significance of CHT described in Section 1 can still be obtained without them. Directedness, on the other hand, is a deal breaker. My contention is that if such a scenario were to arise, then our reasons for believing CHT, or anything like it, would be terminally weakened. Moreover, given the impeccable credentials for the naturalness of Ködel's theory, I think the set theory community would be forced to agree.

But there is another feature of this thought experiment that contrasts with the examples of Section 2 and that provides a little room for pushback. In Section 2, we offered theorems (via Hamkins) that delivered counterexamples to comparability and well-foundedness. In the thought experiment above, we have merely proposed a failure of discovery rather than a proof that there is no large cardinal axiom from which the consistency of Ködel's theory can be established. Given that we are in already in realm of mere fantasy, let us complete our ideal picture and consider what an extra move might look like. Let us suppose that some years after the advent and popularization of Ködel's theory a young logician, Caul Pohen studies the

of NF has been proved. I've opted to leave this little passage as it is, as it provides a pleasing illustration of exactly the kind of event that our fictional account describes.

theory and comes up with a remarkable theorem. We suppose that Pohen shows that there is no large cardinal axiom that we can add to ZFC such that we can start with a model satisfying that large cardinal and then either force or take an inner model to obtain a model of Ködel's theory. Given we have no formal definition of a large cardinal, it is difficult to see how this kind of thing could even be stated as a theorem. However, it is salient to note that there a no currently known examples that fit this template. Every instance of a natural theory that we currently know, even the proper forcing axiom, is consistent relative to a large cardinal axiom. I think it is clear that if Pohen were able to state and prove such a theorem, then the philosophical and foundational significance of the Consistency Hierarchy Thesis would be in tatters.

This is my thought experiment. Of course, it hasn't occurred and we have no good reason to think it will. But I also think the reasons we have to think it won't occur are very weak by conventional mathematical standards. What we do see is that while this ideal situation would be decisive against CHT, it also requires a lot of cards to land a certain way in order for it to take place. In particular, it seems very unlikely that an individual researcher could achieve it since the story has so many moving parts. First, we need a mathematical problem that seems beyond ZFC. Then we need a proposal for solving it that generates interest in the mathematical community. Finally, we need that problem to, at minimum, remain recalcitrant to analysis in the consistency strength hierarchy. Or better, that someone can just show that this theory is not consistent relative to large cardinals. It's a tall order.

3.2. Are we chasing a theorem?

We've now painted a relatively vivid picture emphasizing the challenges involved in attempting to refute the Consistency Hierarchy Thesis. In this section, I would like to offer another explanation as to why a refutation is unlikely to occur. This time, we shall focus shift our focus to the actual rather than counterfactual practice of set theorists and argue that what set theorists seem to expect from relative consistency proofs puts them in – or at least very close to – a position from which delivery of a counterexample is impossible.

3.2.1. Set theorists want more than relative consistency

We start by considering a classic pair of theories where one theory is strictly below the other in consistency strength. We shall then use this example to deliver a generalized version of consistency strength which is provably linear. While the following is standard – if not trivial, it will be hopefully shed a little light and warm up our intuitions.

Theorem 16. *$ZFC + \exists\kappa$ κ is strong $<_{Con}$ $ZFC + \exists\delta$ δ is Woodin, assuming that $ZFC + \exists\kappa$ κ is strong is consistent.*

Proof. Either ZFC with a Woodin cardinal is consistent or it is not. In the latter case, the theorem follows immediately. So we consider the case where it is consistent and fix a model $\mathcal{M}$ witnessing this. We then make the *crucial observation* that if δ is a Woodin cardinal, then there is some $\kappa < \delta$ such that[50]

$$V_\delta \models \kappa \text{ is strong.}$$

[50] A proof of this is essentially delivered in the proof of Proposition 26.13 of (Kanamori, 2003), however, it also follows trivially from the more modern definition of a Woodin cardinals as can be found in Theorem 26.14 of (Kanamori, 2003).

This establishes that ZFC with a Woodin cardinal proves that there is model of ZFC with a strong cardinal, so $V_\delta^{\mathcal{M}}$ thinks there is a strong cardinal, where δ is a Woodin cardinal of $\mathcal{M}$. Thus, we have

$$ZFC + \exists\kappa\ \kappa \text{ is strong} \leq_{Con} ZFC + \exists\delta\ \delta \text{ is Woodin.}$$

To get the strict, $<_{Con}$, we note that if we could prove the consistency of ZFC with a Woodin from the consistency of ZFC with a strong cardinal, then ZFC with a Woodin could prove that it was consistent, contradicting Gödel's second theorem. □

The engine of the proof above lies in its crucial observation; the rest is just metamathematical bookkeeping. But there is also a sense in which the theorem as stated dilutes what the proof actually delivers. If we use the completeness theorem to move from a syntactic setting to a semantic one, then theorem above just tells us that there is a model $\mathcal{M}$ of ZFC that satisfies:

$$\exists\mathcal{N}\ \mathcal{N} \models ZFC + \exists\kappa\ \kappa \text{ is strong} \ \wedge\ \neg\exists\mathcal{P}\ \mathcal{P} \models ZFC + \exists\delta\ \delta \text{ is Woodin.}$$

But this doesn't tell us that the models involved are any good. Are they well-founded? Are they correct? For all we may discern from the statement of the theorem, such $\mathcal{N}$ and $\mathcal{P}$ might disagree on the natural numbers. But the proof tells us more. Before we state this, let us also employ a common set-theoretic idiom and leave the large cardinal assumptions required implicit.[51] Then we have the following.

Theorem 17. *There is a transitive model M of ZFC satisfying that*

$$Con_\beta(ZFC + \exists\kappa\ \kappa \textit{ is strong}) \ \wedge\ \neg Con_\beta(ZFC + \exists\delta\ \delta \textit{ is Woodin}).$$

Proof. Let M be the shortest transitive model with a Woodin cardinal, δ. Then by the crucial assumption from the proof of Theorem 16, we see that V_δ^M thinks there is a strong cardinal. Thus, M satisfies $Con_\beta(ZFC + \exists\kappa\ \kappa \text{ is strong})$. But M cannot satisfy $Con_\beta(ZFC + \exists\delta\ \delta \text{ is Woodin})$ as M was the shortest of its kind. □

Stripped of syntactic detours, this seems to be a more informative theorem. It has an immediate connection to the crucial observation in that it is now essentially a corollary of it. The models employed here all meet a minimum standard of quality in that they are all transitive. I contend that the argument used above is more "set theoretic" in spirit and that it reveals a deeper connection between the theories being compared.

3.2.2. An unattainable target

What happens if we run a little further with the idea of the previous section? So rather than dealing with the syntactic thorniness of mere consistency with all of the counterintuitive features we've seen above, why not work with a collection of models that we have reason to think are good? Any natural set theory should have a well-founded model so why not restrict our attention to the case of well-founded models and their canonical representatives: transitive sets. With that in mind, we might consider a new consistency relation on set theories that better reflects the techniques and attitudes of working set theorists.

[51] In the case below, assuming the existence of a transitive model of ZFC plus a strong cardinal would suffice.

Definition 18. For theories T and S extending ZFC in the language of set theory, let us say that T is β-*consistent relative* to S, abbreviated $T \leq_\beta S$ if

$$ZFC \models_\beta Con_\beta(S) \rightarrow Con_\beta(T).$$

A quick flick back will reveal this is exactly the same as our definition of relative consistency[52] except that we have replaced every use of a consistency or consequence with β-consistency and β-consequence respectively. The underlying idea is that the use of transitive models provides – among other things – a better reflection of the preferences of set theorists. Moreover as we have discussed extensively above, they have good reason for this preference. Ill-founded models do not fit with our intended interpretations of set theory; and worse they often make errors in calculation.[53] Thus, we are investigating what happens when we rid ourselves of this hassle. With this definition in hand, we may then repackage Theorem 17, to obtain

$$ZFC + \exists\kappa\ \kappa \text{ is strong} <_\beta ZFC + \exists\delta\ \delta \text{ is Woodin}.$$

So we might argue that we now have a way of comparing theories that cuts away pathology and gets, at least, a little closer to the information that relative consistency proofs in set theory are able to reveal. I think it is clear that there is often more information still left out of reach in this analysis, but nonetheless, we have a certain kind of improvement on ordinary relative consistency. The obvious question then is: what happens to the Consistency Hierarchy Thesis if we articulate it in the context of β-consistency? The answer may be surprising: it turns it into a theorem!

Theorem 19. *Let T, S be computably axiomatizable theories extending ZFC. Then is must be the case that either: $S \leq_\beta T$ or $T \leq_\beta S$. Moreover, $\leq_\beta$ is a pre-well-ordering on such theories.*

Proof. We just establish the first claim here and leave the latter to an appendix. Nonetheless, the main trick will have already been revealed. Suppose toward a contradiction that T and S provide a counterexample. Then we may fix countable transitive models M_0 and M_1 of ZFC such that:

(i) $M_0 \models Con_\beta(T) \wedge \neg Con_\beta(S)$; and
(ii) $M_1 \models Con_\beta(S) \wedge \neg Con_\beta(T)$.

Using the Lévy-Shoenfield theorem, one can see that, without loss of generality, we may assume that M_0 and M_1 both satisfy $V = L$. And so by the condensation lemma, we may fix $\alpha_0, \alpha_1 < \omega_1$ such that $M_0 = L_{\alpha_0}$ and $M_1 = L_{\alpha_1}$.

Clearly, $\alpha_0 \leq \alpha_1$ or $\alpha_1 \leq \alpha_0$, so suppose the former. Then using (1), we see that there is some transitive $N \in L_{\alpha_0}$ that satisfies T. But then N is also an element of L_{α_1} which contradicts (2). Thus, $\alpha_1 > \alpha_0$. A similar argument then shows that $\alpha_0 > \alpha_1$, which is impossible. □

This simple theorem thus establishes that for any pair of computably axiomatizable theories, their β-consistency strengths are comparable, regardless of whether or not those theories are natural. Given that a natural theory must surely have a transitive model, we seem to have something very like CHT. I think this is quite striking and worthy of note. Of course, the proof is almost absurdly simple; the interest is rather in the set up. So what should we make of this?

[52] See Definition 2. Also see Section 2.2.3. for definitions regarding β-logic.
[53] For example, ill-founded models can think that ZFC is inconsistent.

First of all, $\leq_\beta$ is not $\leq_{Con}$, so we can rightly be accused of moving the goal posts.[54] But does that give us license to simply ignore the result above? I think that would also be too easy. Restricting our attention to transitive models is very much in line with ordinary set theoretic practice. But let us think a little more slowly about this. Recalling that $T \leq_\beta S$ if

$$ZFC \models_\beta Con_\beta(S) \to Con_\beta(T),$$

we see that there are two places where β-logic has entered the story. We are using it *internally* with the Con_β statements and also *externally* we demand that the relationship between the Con_β statements be itself a β-consequence of ZFC. We gave an argument for the internal uses of β-logic in Section 2.2.3., where we noted that in their actual practice theorists prove relative consistency statements using forcing and inner model theory. These arguments are such that if we start with a transitive model, then we also end up with one. Moreover, we have seen many examples above where the pathological behavior of ill-founded models causes counterintuitive and unnatural results. It almost seems obvious that an ill-founded model of set theory cannot be an intended model.

Some of this reasoning also carries over to the external uses of β-logic, but I think the external use is a little more exposed to reasonable doubt. Without it, we are somewhere near the situation of Section 2.2.3., where Hamkins has been able to propose a counterexample albeit without computable axiomatization.[55] With it, the CHT problem is dead. It's just a theorem. So a lot hangs on this move. As has so often been the case in this paper, I don't have anything definitive to say. Worse, I have now moved in the dialectical outsider position, so I don't really expect every reader to come on board. Nonetheless, I'd like to offer a couple of nudges to at least give the reader some pause.

First of all, one might push back against the external use of β-logic by noting that the set of β-logic consequence of ZFC strictly extends ZFC. For example, it is easy to see that $Con(ZFC)$ is β-consequence of ZFC.[56] So perhaps this theory is too strong. In response to this kind of worry, we note that it is easy to see that the β-consequences of ZFC are derivable from a modest, albeit global, large cardinal assumptions.

Proposition 20. *If there is a proper class of inaccessible cardinals, then for all sentences φ of set theory*

$$(ZFC \models_\beta \varphi) \Rightarrow \varphi.$$

Proof. Suppose ψ is true. It suffices to show that there is a transitive model of ZFC in which ψ holds. To see this we observe that since there are unboundedly many inaccessible cardinals, there is a club class of ordinals α such that V_α satisfies ZFC. Thus, using reflection[57] we may obtain such an α where $V_\alpha \models \psi$. □

Informally speaking, this tells us that in the context of a proper class of inaccessible cardinals, any β-consequence of ZFC is already true. Thus, given that ZFC β-consequences hold in

[54] For example, if we let T be ZFC and S be $ZFC + \neg Con(ZFC)$, we have $T \equiv_{Con} S$ and we also have $T <_\beta S$, assuming, say, an inaccessible cardinal.

[55] I believe the question of whether there is non-linearity in the β-consistency hierarchy when we use first order logic externally remains open. The work that seems most closely related to this question occurs in (Aguilera and Pakhomov, 2024). While this paper uses first order logic externally, it also makes use of theories that are not computably axiomatizable. Nonetheless, the proof theoretic techniques utilized there provide a helpful showcase for techniques beyond forcing and inner model theory.

[56] To see this, suppose that M is a transitive model of ZFC. Then clearly $Con(ZFC)$ is true and since every transitive model of ZFC has the correct version of ω, M must think that $Con(ZFC)$ holds.

[57] For a specific formulation, use Theorem IV.7.5 of (Kunen, 2006).

such a modest extension of ZFC, it seems reasonable to take them seriously. Moreover, the conclusion of the proposition is clearly a (very) generalized version of the kind of reflection principle considered by (Feferman, 1991). If we take ZFC seriously, then it seems like the kind of thing we should expect to be true. This doesn't definitively resolve the matter but will hopefully assuage some anxiety. It should also be noted that although Theorem 19 tells us that every pair of natural theories are have comparable β-consistency, it tells us nothing about the direction in which this is resolved or how we might go about figuring that out. As such, it reflects a minor incursion into second order logic where this problem manifests more starkly. In particular, second order logic might be said to decide the continuum hypothesis without giving us any indication of the direction in which it is resolved. Nonetheless, while Theorem 19 tells us nothing about the specific relations between particular theories, it does tell us that, in this rarefied context, the Consistency Hierarchy Thesis has been confirmed.

For a second nudge, it is also worth bearing in mind that what we have called relative β-consistency has periodically appeared in a different guises as a kind of folk ordering on theories. For example, in a footnote of (2014), Steel notes

> There are various ways to attach ordinals to set theories that correspond to the consistency strength order in the case of natural theories. One can look at the provably recursive ordinals, or the minimal ordinal height of a transitive model, for example.

Drawing the last of these out, we might let ρ be a map from the computably axiomatizable theories to the ordinals, such that: $\rho(T) = \alpha$ if α is least such that there is a transitive model M satisfying T where $\alpha = Ord^M$. From this, we can then obtain a relatively simple pre-well-ordering, $\preceq_{rank}$ on computably axiomatizable theories. I think it's fair to say that this definition has the feel of folklore and that it clearly puts everything we want in a sensible order. If you wanted to put theories into a pre-well-order, it's probably the most obvious way to do it. However, I think the philosophical and foundational significance of this ordering is less obvious. While it certainly works in the mathematical sense of providing an ordering, it's not so clear how this relates to CHT. It turns out that this obvious pre-well-ordering is generally equivalent to the β-consistency ordering.

Proposition 21. *Suppose T and S are computably axiomatizable theories in $\mathcal{L}_\in$ that extend ZFC, then*[58]

$$T \preceq_{rank} S \Leftrightarrow T \leq_\beta S.$$

Again, this is hardly definitive evidence that relative β-consistency is the right way to go. But the fact that there are at least a couple of ways one might bump into this ordering suggests that something significant is occurring here.

So much for my nudges. Where does this leave us? Since we've changed the subject, we certainly haven't confirmed CHT. Nonetheless, I do think we can take a couple of lessons away from this experience. The first of these is that Theorem 19 highlights the fact that ordinary set theoretic methods cannot be used to obtain a counterexample to CHT. A successful refutation of CHT would require, for example, a pair of theories with regard to which we *cannot prove* that

[58] We save the proof for the appendix.

one theory is consistent relative to the other. As such, it requires a proof that there is no proof; i.e., a consistency proof is required. But in contrast to the techniques of ordinary set theory, such consistency proof cannot be obtained by forcing or inner model theory. As we have seen, these proofs take us from transitive models to transitive models and are thus, hostage to Theorem 19 where no counterexamples are available. Of course, the fact that techniques from outside the usual ken of set theory – like computability theory and proof theory – may be required to address this problem has no impact on CHT as a mathematical problem. But – and this brings us to the second lesson – it may have some affect on how we evaluate the foundational significance of the problem. If the only way to obtain a counterexample is to work within the pathological realms of nonstandard models, then perhaps we don't have to take a counterexample to CHT very seriously at all. Indeed and more provocatively, perhaps a pair of theories providing a counterexample to CHT is – in itself – evidence that at least one of those theories isn't natural.

3.3. Making sense of it all

Let's close this paper by trying to pull its many threads together. We started by defining what it means for one theory to be consistent relative to another and from there we were able to describe what we called the Consistency Hierarchy Thesis. We were able to get an understanding of what makes consistency strength interesting from a mathematical perspective. We then noted that CHT provides a tantalizing answer to the question of what lies beyond ZFC: a surprising amount of order. Moreover, we noted the fundamental role that CHT plays in our arguments for the existence of large cardinals and our understanding of set theory's influence on ordinary mathematics. Nonetheless, we also observed that CHT is not a genuine mathematical problem. In particular, the notion of "natural theory" has no mathematically precise formulation and yet it plays a crucial role in the articulation of CHT. As such, we started to wonder about the prospects for refuting CHT. With this in mind, we then explored a number of putative counterexamples proposed by Joel David Hamkins. As we suspected, in each of these cases, the claim for the naturalness of the theories involved was exposed and vulnerable to pushback. So despite the impressive ingenuity inherent in each of these examples, we were not optimistic that the broader set theory community would embrace them as successful. We did not, however, find this to be a cause for simple celebration on behalf of CHT. Rather, we worried about the dialectical advantage possessed by those aiming to defend CHT against putative counterexamples. The bar for naturalness among theories is set so very high. This then prompted a different tack. Instead of considering real examples, we asked what an ideal counterexample to CHT might look like. We ended up with a fictional account that – were it to occur – would likely succeed in refuting CHT and gain acceptance from the mathematical community that it had. But the story had a long wish list. A number of clever people had to do a number of brilliant things over a probable period of decades. Nothing like it has occurred and nothing like it seems to be in the process of occurring right now. This prompted our final shift in perspective. Rather than taking CHT literally, we considered the kinds of argument that set theorists use in establishing relative consistency claims. In particular, forcing and inner model theory always take us from one transitive model to another. When we reformulated CHT by immersing it in the context of transitive models, our problem simply disappeared since CHT became a theorem.

We seem to have painted a picture in which CHT appears almost immune to refutation. At the least, we've seen that a successful counterexample will be quite strange in that the standard

techniques of forcing and inner model theory cannot deliver it. Perhaps the strangeness of the tools used to establish such a result would – while delivering a minor mathematical miracle – only offer us philosophical crumbs in return. Part of appeal of CHT lies in the fact that the interpretations yielding its instances preserve a great deal of meaningful structure. If this were absent, then the foundational significance of such an example would be difficult to evaluate. For reasons like these, we may end up in a situation where any proposed counterexample to CHT is rejected since theories that are not connected by forcing or inner model theory are – almost by fiat – unnatural. While we are not currently in such a situation, this could seem unappealing or even question begging. But I think it actually gets us closer to the right way to understand CHT. As we suggested above, it bears an interesting relationship with the Church-Turing thesis, although the CHT story is far less complete. However, if natural theories continue to be regimented by the large cardinal hierarchy using forcing and inner model theory, then we may well come to see things from the other side of the table: we might see natural theories as being – in an informal sense – exactly those that can be accessed from large cardinal axioms by forcing and inner models. There would be a long way to go along that road before we get to such a position, but it is important to remember that – if this road were to be traveled – we would not only have a longer list of elegant equiconsistency proofs, the tools developed in producing those proofs would also give us insight into what is making those theories seem natural. The point I'm trying to make is that this would not be a victory via mere enumeration, greater understanding will also need to come along for the ride. In this spirit, I'd like to end this paper with a shoutout to two projects that show great promise in developing such understanding and which go far beyond the relatively superficial analysis of this paper. The longest running program in this area takes place within inner model theory. John Steel describes three closely related hierarchies: the consistency strength hierarchy of this paper; the Wadge hierarchy on homogeneously Suslin sets of reals; and the mouse order on canonical inner models for large cardinals. For Steel, the final ordering on mice provides the foundation for the other two (Steel, 2024).[59] Thus, the consistency strength hierarchy is explained through a deeper hierarchy of models that instantiate natural theories. More recently, James Walsh has taken a different angle on this problem of understanding by asking very directly: what makes a theory natural? This work draws on techniques from ordinal analysis in proof theory and analogies with Martin's Conjecture in computability theory (Montalbán and Walsh, 2019; Walsh, 2025). While a lot of work remains to be done and many deep questions remain open, the progress toward understanding is plain to see.

Acknowledgements. I'd like to thank John Steel for some helpful discussions about this material. I'd also like to thank Pen Maddy for some helpful comments on an earlier version of this paper. Thanks also to the Irvine LPS Logic Seminar for letting me present this material. And thanks to the anonymous referees who helped me improve this paper.

References

Aguilera, J. P. and Pakhomov, F. (2024). Non-linearities in the analytical hierarchy. Preprint.

Black, R. (2000). Proving church's thesis. *Philosophia Mathematica*, 8(3):244–258. DOI: https://doi.org/10.1093/philmat/8.3.244.

[59] For an overview on the state of the art on this topic see the introduction to (Steel, 2022).

Boolos, G. (1984). The logic of provability. *The American Mathematical Monthly*, 91:470–480. DOI: https://doi.org/10.1080/00029890.1984.11971467.
Cohen, P. (1963). The independence of the continuum hypothesis. *Proceedings of the National Academy of Sciences of the USA*, 50(6):1143–1148.
Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, 56(1):1–49. DOI: https://doi.org/10.2307/2274902.
Feferman, S., Friedman, H. M., Maddy, P., and Steel, J. R. (2000). Does mathematics need new axioms? *The Bulletin of Symbolic Logic*, 6(4):401–446. DOI: https://doi.org/10.2307/420965.
Gödel, K. (1940). *Consistency of the Continuum Hypothesis*. Princeton University Press.
Gödel, K. (1986). On formally undecidable propositions of principia mathematica and related systems. In Feferman, S., editor, *Kurt Gödel Collected Works Volume I: Publications 1929-1936*, volume 1. Oxford University Press, New York. DOI: https://doi.org/10.1093/oso/9780195147209.003.0014.
Goldberg, G. (2022). *The Ultrapower Axiom*. De Gruyter, Berlin, Boston. DOI: https://doi.org/10.1515/9783110719734.
Grotenhuis, L. (2022). Natural axiomatic theories and consistency strength: A lakatosian approach to the linearity conjecture. Report.
Hamkins, J. D. (2025). Nonlinearity and illfoundedness in the hierarchy of large cardinal strength. *Monatshefte für Mathematik*. DOI: https://doi.org/10.1007/s00605-025-02082-1.
Holmes, M. R. and Wilshaw, S. (2024). Nf is consistent. https://arxiv.org/abs/1503.01406.
Kanamori, A. (2003). *The Higher Infinite: Large Cardinals in Set Theory from Their Beginnings*. Springer. DOI: https://doi.org/10.1007/978-3-540-88867-3.
Kaye, R. and Wong, T. L. (2007). On interpretations of arithmetic and set theory. *Notre Dame Journal of Formal Logic*, 48(4):497–510. DOI: https://doi.org/10.1305/ndjfl/1193667707.
Kechris, A. S. (1995). *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. Springer. DOI: https://doi.org/10.1007/978-1-4612-4190-4.
Kunen, K. (2006). *Set Theory: An Introduction to Independence Proofs*. Elsevier, Sydney.
Kunen, K. (2009). *The Foundations of Mathematics*. Mathematical Logic and Foundations. College Publications.
Lewis, A. (1998). Large cardinals and large dilators. *The Journal of Symbolic Logic*, 63(4):1496–1510. DOI: https://doi.org/10.2307/2586663.
Lindström, P. (2003). *Aspects of Incompleteness: Lecture Notes in Logic 10*. Lecture Notes in Logic. Taylor & Francis.
Maddy, P. (2011). *Defending the Axioms: On the Philosophical Foundations of Set Theory*. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/acprof:oso/9780199596188.001.0001.
Maddy, P. and Meadows, T. (2020). A reconstruction of steel's multiverse project. *Bulletin of Symbolic Logic*, 26(2):118–169. DOI: https://doi.org/10.1017/bsl.2020.5.
Martin, D. A. (1998). Mathematical evidence. In Dales, H. G. and Oliveri, G., editors, *Truth in Mathematics*. Clarendon Press. DOI: https://doi.org/10.1093/oso/9780198514763.003.0012.
Martin, D. A. (n.d.). *Determinacy*. Unpublished book manuscript.
Meadows, T. (2021). Two arguments against the generic multiverse. *Review of Symbolic Logic*, 14(2):347–379. DOI: https://doi.org/10.1017/S1755020319000327.
Montalbán, A. and Walsh, J. (2019). On the inevitability of the consistency operator. *Journal of Symbolic Logic*, 84(1):205–225. DOI: https://doi.org/10.1017/jsl.2018.65.
Steel, J. (2024). The comparison lemma: Young set theory workshop 2024.

Steel, J. R. (2014). Gödel's program. In Kennedy, J., editor, *Interpreting Gödel: Critical Essays*, pages 153–179. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511756306.012.

Steel, J. R. (2022). *A Comparison Process for Mouse Pairs*. Lecture Notes in Logic. Cambridge University Press. DOI: https://doi.org/10.1017/9781108886840.

Todorcevic, S. (2014). *Notes on Forcing Axioms*. World Scientific. Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore: Volume 26. Edited by Chitat Chong, Feng Qi, Theodore A. Slaman, W. Hugh Woodin and Yue Yang. DOI: https://doi.org/10.1142/9013.

Walsh, J. (2025). On the hierarchy of natural theories. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2106.05794.

Woodin, W. (2001). The continuum hypothesis, part II. *Notices of the AMS*, 48(7):681–690.

A. Appendix

We show here – among other things – that the $\leq_\beta$ ordering is the same as the ordering obtained by considering the ordinals of the least transitive model of a theory. We start with a simple technical fact.

Lemma 22. *Suppose $x \in L_\alpha \cap \mathbb{R}$ and β is the second admissible ordinal greater than α. Then if $\exists y \in \mathbb{R}\ \varphi(y, x)$ is true statement where $\varphi(y, x)$ is Σ^0_1, then there is some $y \in L_\beta$ such that $\varphi(y, x)$.*

Proof. First note that there is tree T on ω^2 with $T \in L_\alpha$ such that for all $y \in \mathbb{R}$

$$\varphi(y, x) \Leftrightarrow y \in [T(x)].$$

Then, if γ is the first admissible above α, $T(x)$ is ill-founded iff L_γ thinks there is no order preserving map from $T(x, y)$ into the ordinals. And then this holds iff L_β thinks there is a branch $y \in \mathbb{R}$ through $T(x)$. Thus, there is some $y \in L_\beta$ such that $\varphi(x, y)$. □

The following lemma establishes that if T has a transitive model, then it has a model of minimal height in L.

Lemma 23. *Suppose T is a computable theory with a transitive model and that $\alpha = Ord^M$ where M is the shortest of these models. Then T has a model N in L with $Ord^N = \alpha$. In fact, such a model will be an element of L_β for any $\beta > \alpha$ where $L_\beta \models ZFC^-$.*

Proof. Let T and α be as described and let M be a transitive model of T with $Ord^M = \alpha$. Clearly $\alpha < \omega_1$ so we may fix $x_\alpha \in \mathbb{R}$ that codes a well-ordering of length α. First we claim that such an x_α exists in L. To see this suppose not. Then we must have $\omega_1^L < \alpha < \omega_1$. Now note that the statement

> There is countable transitive model of T.

is Σ^1_2 and so is true in L by the Lévy-Shoenfield theorem. Thus, we may fix a countable transitive model M^* of T with $M^* \in L$ and $|M^*|^L < \omega_1^L$. But then $Ord^{M^*} < \omega_1^L < \alpha$ contradicting the minimality of α.

Now with $x_\alpha \in \mathbb{R} \cap L$ coding α in hand, we observe that our assumptions imply that:

> There is some $m \in \mathbb{R}$ coding a model satisfying T and such that the ordinals of that model are isomorphic to the well-ordering encoded by x_α.

This statement is Σ^1_1 in x. Thus using Lemma 22, we may obtain some $m \in L_\beta \cap \mathbb{R}$ witnessing its truth, where $\beta > \alpha$ and $L_\beta \models ZFC^-$. The real m can then be decoded and collapsed in L_β to obtain a model N with $Ord^N = \alpha$. □

We now formalize the definition of our ranking function on theories.

Definition 24. Let ρ be a function from a computably axiomatizable theories to the ordinals be such that $\rho(T) = \alpha$ if α is the least α such that there is a transitive model M of T with $Ord^M = \alpha$; and let $\rho(T) = *$ if there is no such model, where $*$ is some non-ordinal from V_ω. Let us say that $T \preceq_{rank} S$ if $\rho(T) \leq_* \rho(S)$, where $\leq_* \subseteq (Ord \cup \{*\})^2$ is the usual ordering on the ordinals with $*$ added to the end.

Informally speaking, when we say $T \preceq_{rank} S$, we are saying that the shortest model of T, if there is one, is shorter than the shortest model of S.

Theorem 25. *$T \preceq_{rank} S$ is Π^1_2.*

Proof. It is easy to see that Lemma 23 implies $T \preceq_{rank} S$ iff

> For all $\alpha < \omega_1$, if $L_\alpha \models ZFC^-$, then $L_\alpha \models T \preceq_{rank} S$.

Then since this statement is Π^1_2, we are done.

□

Theorem 26. *If $T \preceq_{rank} S$, then $T \leq_\beta S$.*

Proof. Suppose $T \not\leq_\beta S$. Then there is a transitive model M of ZFC^- satisfying $Con_\beta(T)$ but not $Con_\beta(S)$. Let $\alpha = Ord^M$. Then by Lemma 23, we see that L_α thinks that $\rho(T) < \alpha$ and $\rho(S) = *$. Thus, $L_\alpha \models T \preceq_{rank} S$, which suffices by the proof of Theorem 25. □

We then note that the converse can fail if we consider theories that don't extend ZFC. Here is a somewhat artificial example demonstrating this.

Proposition 27. *There are computably axiomatizable T and S such that $T \leq_\beta S$ but $T \not\preceq_{rank} S$, supposing there is a transitive model of ZFC.*

Proof. Let S be ZFC and let T be KP plus the statement that there is a transitive model of ZFC. Let L_α be such that α is least such that L_α satisfies ZFC. Let L_β be the next admissible ordinal after α. Then β is least such that $L_\beta \models T$. So clearly we have $S \preceq_{rank} T$ and $T \not\preceq_{rank} S$. We now show that $T \leq_\beta S$. To do this, first let γ be the least ordinal greater than β satisfying ZFC^-. Note that $\gamma > \beta > \alpha$. Then let N be an arbitrary model of ZFC^- and suppose that N satisfies $Con_\beta(S)$. Then we must have $L_\alpha \in N$ and indeed $L_\gamma \in N$. Thus, N also satisfies $Con_\beta(T)$ as required. □

However, if we restrict our attention to theories extending ZFC in $\mathcal{L}_\in$ this glitch disappears.

Problem A.1. Suppose T and S are computably axiomatizable theories in $\mathcal{L}_\in$ that extend ZFC, then

$$T \preceq_{rank} S \Leftrightarrow T \leq_\beta S.$$

Proof. We just need the $\Leftarrow$ direction here, so suppose $T \not\preceq_{rank} S$. Then $\rho(T) >_* \rho(S)$ and so S must have a transitive model. If there are no transitive models of T, we trivially get $T \not\leq_\beta S$. So suppose that T also has a transitive model. Using Lemma 23, we may fix M satisfying S with $Ord^M = \rho(S)$ and fix β least such that L_β satisfies ZFC^- and $M \in L_\beta$. Similarly, we may fix N satisfying T with $Ord^N = \rho(T) > \rho(S)$. Now suppose toward a contradiction that $Ord^N < \beta$. But then $L^N = L_\gamma$ for some $\gamma < \beta$ and by Lemma 23, we must have $M \in L_\gamma$. But since L_γ satisfies ZFC this means β cannot have been the least model of ZFC^- with $M \in L_\beta$, so we have a contradiction. □

Journal for the Philosophy of Mathematics

OPEN ACCESS

Citation: SHAPIRO, Stewart (2025). Semantic indeterminacy, concept sharpening, and set theories. *Journal for the Philosophy of Mathematics*. 2: 143-160. doi: 10.36253/jpm-3888

Received: July 16, 2024

Accepted: December 12, 2024

Published: December 30, 2025

ORCiD
SS: 0000-0002-4895-0772

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Firenze University Press
www.fupress.com

Semantic indeterminacy, concept sharpening, and set theories

STEWART SHAPIRO

Ohio State University & University of Connecticut, USA.
Email: shapiro.4@osu.edu

Abstract: Friedrich Waismann once suggested that mathematical concepts are not subject to open-texture; they are "closed". This is not quite right, as there are some traditional mathematical notions that were, at least at one time, open-textured. One of them is the notion of "polyhedron" following the history sketched in Imre Lakatos's *Proofs and refutations*. Another is "computability", which has now been sharpened into an arguably closed notion, via the Church-Turing thesis.

There are also some mathematical notions that have longstanding, intuitive principles underlying them, principles that later proved to be inconsistent with each other, sometimes when the notion is applied to cases not considered previously (in which case it is perhaps an instance of open-texture). One example is "same size", which is or was governed by the part-whole principle (one of Euclid's Common Notions) and the one-one principle, now called "Hume's Principle". Another is the notion of continuity.

The purpose of this paper is to explore the notion of "set" and other related notions like "class", "totality", and the like. I tentatively put forward a thesis that this notion, too, is or at least was subject to open-texture (or something like it) and has been sharpened in various ways.

This raises some questions concerning what the purposes of a (sharpened) theory of sets are to be. And questions about how one goes about trying to give non-ad-hoc explanations or answers to various questions.

Keywords: Open texture, Waismann, Semantic indeterminacy.

1. Introduction: open texture and mathematics

Friedrich Waismann (1945) introduced the term "open texture" into philosophy. As a first approximation, we might say that a predicate P, from a natural language, exhibits *open texture* if it is possible for there to be an object a such that nothing concerning the established use of P, and nothing concerning the non-linguistic facts, determines that P holds of a, nor does anything determine that P fails to hold of a.[1]

[1] Here I primarily take open texture to hold (or not) among words and phrases, but one can also think of concepts as being open textured (or not). I often use a term like "notion" meant to be ambiguous between phrases and concepts.

This characterization may not be quite right, as it does not seem to distinguish instances of open texture from borderline cases of a vague predicate, say in a Sorites series—at least according to some views of vagueness (e.g., Shapiro, 2006b). With open texture, the cases typically come out of the blue, sometimes as the result of philosophical thought experiments and sometimes in the normal use of the expressions, but in new contexts. The examples speak for themselves.[2]

In effect, if a predicate P is open-textured, then the truth of some sentences in the form Pa is left open by the use of the language and the non-linguistic facts: nothing languages users have said or done to date—whether by way of the ordinary use of the term in communication or in an attempt to stipulate its meaning—fixes how the term should be applied to the new cases. Open texture is a kind of semantic indeterminacy.

Simon Blackburn's *Oxford Dictionary of Philosophy* (1994) contains the following entry:

> **open texture**: The term, due to Waismann, for the fact that however tightly we think we define an expression, there always remains a set of (possibly remote) possibilities under which there would be no right answer to the question of whether it applies ...
> For example, [the open texture] of the term "mother" ... is revealed if through technological advance differences open up between the mother that produces the ovum, the mother that carries the foetus to term, and the mother that rears the baby. It will then be fruitless to pursue the question of which is the 'real' mother, since the term is not adapted to giving a decision in the new circumstances.

Waismann himself says that mathematical terms—presumably all mathematical terms—are *not* subject to open texture; they are "closed":

> If, in geometry, I describe a triangle, e.g., by giving its three sides, the description is *complete*: nothing can be added to it that is not included in, or at variance with, the data. (Waismann, 1945, 125)

To echo Waismann's somewhat Wittgensteinian rhetoric, however, what should we say of various three-sided figures in non-Euclidean spaces, say those with variable curvature, where it is not clear what counts as a "straight line"?

Perhaps "Euclidean triangle" is a better candidate for a term that has no open texture. When the restriction to Euclidean spaces is made explicit, the case for the term being "closed" is better. At least it is not as easy to think of cases that would or even might indicate the open texture of this notion.

Shapiro (2006a) argues that "polyhedron" and "sameness of area" were, at least once upon a time, subject to something like open texture, and yet these notions are mathematical, if any are. Theorems about polyhedra and area are found in mathematics texts from antiquity, long before any rigorous formal definitions for these notions were provided.

[2] It is fair to say that "open texture" is homological—it is itself subject to open texture. It turns out that just about all empirical terms from ordinary language are subject to open texture, as are at least some technical terms that arise in the development of the sciences. There is no reason to think that technical terms in philosophy are any different.

To be sure, each of these terms, and many others, were eventually sharpened via rigorous definitions, sharpened to the point where it at least appears that there is no remaining open texture.

2. Proofs and refutations

Let us briefly revisit one such case here, highlighted by Imre Lakatos *Proofs and refutations*" (1976). The bulk of that little book is a lively dialogue involving a class of rather exceptional mathematics students. The dialogue is a rational reconstruction of the history of what may be called "Euler's theorem or, perhaps "Euler's conjecture".[3]

> Consider any polyhedron. Let V be the number its vertices, E the number of its edges, and F the number of its faces. Then $V - E + F = 2$.

The reader is first invited to check that the equation holds for standard polyhedra, such as rectangular solids, pyramids, tetrahedra, icosahedra, and dodecahedra. One would naturally like this quasi-inductive "evidence" confirmed with a proof, or else refuted by a counterexample. As indicated by the title of the book, Lakatos provides examples of both.

The dialogue opens with the teacher presenting a proof of Euler's conjecture. Think of a given polyhedron as hollow, with its surface made of thin rubber. Remove one face and stretch the remaining figure onto a flat wall. Then add lines to triangulate all of the polygonal faces, noting that in doing so we do not change $V - E + F$. When the figure is fully triangulated, start removing the lines one or two at a time, doing so in a way that does not alter $V - E + F$. At the end, we are left with a single triangle, which, of course, has 3 vertices, 3 edges, and 1 face. So for that figure, $V - E + F = 1$. If we add back the face we removed at the start, we find that $V - E + F = 2$ for the original polyhedron. QED.

The class then considers a series of counterexamples to Euler's conjecture. These include a picture frame, a cube with a cube-shaped hole in one of its faces, a cube with cube-shaped hole in its interior, and a "star polyhedron", a figure with pyramids sticking out from some of its sides.

A careful examination shows that each counterexample violates (or falsifies) at least one of what Lakatos calls the "hidden lemmas" of the teacher's proof. In some cases, the three-dimensional figure in question cannot be stretched flat onto a surface after the removal of a face (at least not without some of the faces overlapping others). In other cases, the stretched plane figure cannot be triangulated without changing the value of $V - E + F$ (or cannot be triangulated at all), and in still other cases, the triangulated figure cannot be decomposed without altering the value of $V - E + F$.

The dialogue then takes an interesting turn, especially given that it more or less follows some threads in the history of mathematics, or at least the history of this mathematics. Some students declare that the counterexamples are what Lakatos calls "monsters" and do not refute Euler's conjecture. The idea here is to insist that the weird figures in question are not really polyhedra, or are not really proper polyhedra. One can imagine some philosophers arguing that an analysis of the concept of "polyhedron" reveals this.

[3]Or at least it could be called that if there were not hundreds of other results that equally deserve the title. Parts of the actual history of this particular theorem or conjecture are sketched in Lakatos's footnotes.

The class does consider a desperate attempt along those lines: one *defines* a polyhedron to be a figure that can be stretched onto a surface once a face is removed, and then triangulated and decomposed in a certain way. That would make the teacher's "proof" into a stipulative definition of the word "polyhedron". This move is quickly dismissed.

A second attempt to resolve the matter is to restrict the theorem so that the proof holds: the proper theorem is that for any convex, "simple" polyhedron, $V - E + F = 2$. Apparently, an advocate of this maneuver is content to ignore the interesting fact that $V - E + F = 2$ does hold for some concave, and some non-simple polyhedra.

A third line is to take the counterexamples to refute Euler's conjecture, and to declare that the notion of "polyhedron" is too complex and unorderly for decent mathematical treatment. Apparently, those inclined this way just lose interest in the notion, at least in their capacity as mathematicians.

A fourth maneuver is to accept the counterexamples as refuting Euler's conjecture, and then look for a generalization that covers the Eulerian and non-Eulerian polyhedra in a single theorem.[4]

It is straightforward to interpret the situation in Lakatos's dialogue—or, better, the history it reconstructs—in terms of Waismann's account of language. The start of the dialogue refers to a period in which the notion of polyhedron had an established use in the mathematical community (or communities). As noted, theorems about polyhedra go back a long way in mathematics. Nevertheless, the word, or notion, or concept, was not defined in a way that either explicitly included or explicitly ruled out the alleged counter-examples. Euclid, for example, defines a "solid" to be "that which has length, breadth, and depth" (Book XI, Definition 1), but he makes no mathematical use of that definition. In that respect, it is similar to some of the definitions in Book I, like "a point is that which has no part". Surely, a necessary condition for a solid to be a polyhedron is that it be bounded by plane polygons (i.e. networks of edges all of which lie in the same plane and enclose a single area). But what are the sufficient conditions for being a polyhedron?

The mathematicians of antiquity were working with a notion governed more by a Wittgensteinian family resemblance than by a rigorous definition that determines every case one way or the other. In other words, the notion of polyhedron exhibited open texture.

Note that this open texture did not prevent mathematicians from working with the notion, and proving things about polyhedra—witness, again, Euclid's *Elements*. Still, at the time, it simply was not determinate whether a picture frame counts as a polyhedron. The same goes for a cube with a cube-shaped hole in one of the faces, etc.

When the troubling cases did come up, and threatened to undermine a lovely generalization discovered by the great Euler, a decision had to be made. As Lakatos shows, different decisions were made, or at least proposed. Those who found the teacher's proof compelling (at least initially) could look to its details—to what Lakatos calls its "hidden lemmas"—to shed light on just *what a polyhedron is*. And those who found the counterexamples compelling (but do not lose interest in the notion) can look to the details of the proof, and to the counterexamples, to formulate a more general definition of "polyhedron", in order to find the characteristics that make some polyhedra, and not others, Eulerian (i.e., such that $V + E - F = 2$).

[4]Toward the end of the book (in a late section possibly added by Lakatos's students and editors), one character proposes what we may call an algebraic definition: a polyhedron just is a collection of things called "vertices", "edges", and "faces" that have certain relations to each other.

In this case, at least, both approaches proved fruitful. We can look back on the history and see how much was learned about the geometry of (Euclidean) space.

3. Computability

The history of mathematics provides many similar examples. Consider, for example, the notion of a real-valued function. Mathematicians were confronted with "monsters", such as functions that are discontinuous everywhere, and functions that are continuous everywhere, but differentiable nowhere.[5] Or the notion of area (or integral), continuous function, convergent series, derivative (see Smith, 2015), etc. Eventually, all of these notions received rigorous definitions, to the point where it at least appears that no open texture remains.

Shapiro (2006a) argues that, at least in the early 1930's, the notion of computability was subject to open texture. At some point, and in many contexts today (mathematical and otherwise), matters of feasibility are relevant to what counts as computable. For example, an Ackermann function is not computable in any reasonable or at least any practical sense, since it would take more particles in the universe to compute its value for even small inputs (like the pair $\langle 5, 5\rangle$). Some of the relatively early debates over the Church-Turing thesis saw key mathematicians and mathematical logicians claiming that certain recursive functions are not computable due to matters of feasibility (Mendelson, 1963). That at least suggests that it was not completely clear at the time whether matters of feasibility should matter for computability.

The mathematical (and philosophical) field of computability gained considerable insight and depth when partial functions—functions that are not defined on every input—were considered. Of course, one always knew that division by the natural numbers is not defined on 0, and that many natural numbers do not have square roots, but the consideration of partial functions in computability goes far beyond these easy examples. It is not much of an exaggeration to say that the entire enterprise of recursive function theory is built around them.

For example, one can enumerate all of the Turing machines, and thus one can enumerate (with repetitions) all partial recursive functions. However, one cannot "diagonalize" out of this enumeration, just because (some of) the functions are partial.

On the contemporary definitions, a partial function f on the natural numbers is said to be computable if and only if there is an algorithm A such that, for each natural number n, if A is given n as input:

- If f is defined on n, then A produces $f(n)$, and
- if f is not defined on n, then A produces no output at all; typically, A runs forever in these cases.

Oliver and Smiley (2013, pp. 24-326), argue that this definition is incorrect. Their conclusion seems to be based on intuitions about what it is to be computable; they seem to attempt a traditional conceptual analysis of "partial computability" (i.e., computability of partial functions):

> So what are the computable functions? They are those for which there is an algorithm that delivers the value whenever there is one, and registers that there is no value when there is none, whether implicitly by halting without producing

[5] See Youschkevitch (1976), or any text in the history of mathematics.

> an approved numerical output, or explicitly by adding O to approved outputs and halting with O as output. (p. 325)

I submit that the intuitive, pre-theoretic notion of computability, available in the 1930's and for a bit after, did not adjudicate between the contemporary conception of what may be called "partial computability" and the one insisted on by Oliver and Smiley. This is, or was, a case of open texture. Nothing that mathematicians said or did, either by way of how the notion of computability was were used, nor in attempts at a definition, determined what to say about the computability of partial functions. In retrospect, the contemporary account—the one criticized by Oliver and Smiley—proved to be fruitful.

One can ask whether the word "computable", or the corresponding concept, has the same meaning today as it did in the 1930s when computability became a focus of the mathematical and logical worlds. In the third item in the Analytic-Synthetic series, Waismann asked if the meaning of the word "time" was changed when it was discovered how to measure time or, perhaps better, when it was discovered that there are reliable ways to measure time. He wrote:

> whether the meaning of "time" ... changes when a method of measuring is introduced, we were thinking of the meaning of a word as clear-cut. What we were not aware of was that there are no precise rules governing the use of words like "time" ... and that consequently to speak of the "meaning" of a word and to ask whether it has, or has not changed in meaning, is to operate with too blurred an expression. (Waismann, 1951, p. 53)

The "too blurred expression" here is "means the same as", when applied to words and concepts separated by sufficient periods of time (or place).

In the case at hand, there is no sharp fact of the matter whether the notion of computability in play before the 1930's is the same or different from the one invoked today, especially as to how that term is supposed to apply to partial functions.

Something similar can be said about other notions mentioned above: polyhedron, sameness of area, and the like. As noted, all of these notions eventually received rigorous definitions which at least appear to be free of open texture. But they did not start out with such definitions, even though there was a rich mathematical practice around them. Following Waismann, to ask whether the rigorous definitions got things right with respect to previous practice is to operate with too blurred an expression. Mathematics does resist open texture, but is not immune to it (cf. Shapiro and Roberts, 2021). Indeed, one might say that, in mathematics, open texture is more the rule than the exception, at least historically.

4. Contradictory principles

Occasionally, there are some principles associated with a given notion. Some of these are thought to be constitutive of the notion, telling us just what the notion is. And sometimes some of these principles are found to be inconsistent with each other. This may happen when the notion is applied to heretofore unconsidered cases, thus invoking something at least in the neighborhood of open texture. The difference is that with open texture, the application of the notion to the new cases is indeterminate—nothing speakers have said or done determines

whether the notion applies or not. In the present cases, the principles conflict with each other (dialethism aside).

Although matters are controversial (aren't they always?), perhaps truth is one such notion. The principles in question are the two directions of the so-called "T-scheme":

- If 'φ' is true, then φ.
- If φ then 'φ' is true.

A Liar sentence or proposition would be a heretofore unconsidered case.

Here we will stick to mathematical notions. Perhaps the clearest case is the relation of two collections being the same size. One arguably constitutive principle of "same size" is the *Part-Whole Principle*, enshrined in one of Euclid's Common Notions:

> The part is smaller than the whole.

The other is now called *Hume's Principle*:

> Two collections are the same size if and only if there is a one-to-one correspondence from one onto the other.

Of course, these come into conflict if the collections are infinite.

Beginning with Aristotle, philosophers argued that there are no completed infinite totalities. Sometimes this was defended on the grounds of the Part-Whole principle. What we today call "Hilbert's Hotel", a feature of infinite collections, was called "Galileo's Paradox", a refutation of the very idea of a completed infinite collection. When the calculus was being developed, mathematicians like Galileo and Leibniz came to accept completed infinite totalities, but insisted that they don't have sizes, or that they cannot be compared for size.[6]

Of course, the contemporary Cantorian resolution is to simply abandon the Part-Whole Principle, in favor of what is now called Hume's Principle. Someone who favors contemporary practice may suggest that the Part-Whole Principle is not, *and perhaps never was*, constitutive of the notion of "same size". Adopting that principle is a (perhaps understandable) mistake that our ancestors made, due to not considering enough cases.

In "Measuring the size of infinite collections of natural numbers: was Cantor's theory of infinite number inevitable?", Paolo Mancosu (2009) shows how it is indeed possible to develop a consistent theory of "same size" based on the Part-Whole Principle, rejecting Hume's Principle. The account is not nearly as elegant as the Cantorian one, and requires some rather ad hoc parameters, but it is consistent.

Another example (perhaps) is the notion of a continuous substance. One longstanding principle is that a continuous substance has a kind of unity. There is something that binds it together, making it a coherent whole. This goes back to Aristotle's *Physics* (227a6):

> The continuous is just what is contiguous, but I say that a thing is continuous when the extremities of each at which they are in contact become one and the same and are (as the name implies) contained in each other. Continuity is impossible if these

[6] See the papers in Hellman and Shapiro (2021).

> extremities are two. This definition makes it plain that continuity belongs to things that naturally, in virtue of their mutual contact form a unity.

Let us call this (alleged) feature of continuous substances *viscosity*.

When it comes to mathematical things like lines, one fallout of viscosity is sometimes called *indecomposibility*: it is impossible to cleanly break a continuous substance into two parts. For Aristotle, if one attempts to break a line, for example, one will produce something new, or, perhaps better, one will make actual something that was only potential before, namely the boundaries of the new lines (the endpoints).

In contemporary intuitionistic analysis and in smooth infinitesimal analysis, also based on intuitionistic logic, viscosity plays out differently: for any given line a is not the case that there are two lines b, c that are both part of a, discrete from each other, and such that b and c together are a. It is sometimes put as a metaphor: if you try to cut a line, something will stick to the knife.

A second intuitive principle about continuous lines is expressed by the intermediate value theorem:

> If two lines cross each other, then they intersect: there is at a point common to the two lines where they meet.

This principle is also displayed, or perhaps presupposed, in one of the so-called "gaps" in Euclid's *Elements*. It occurs in the very first proposition (I:1): Euclid constructs two circles that cross each other, and instructs the reader to let A be the point where the circles meet. One might ask how one is supposed to know that there is a point where the two circles meet. Perhaps this is implicit in the continuity of the lines.

In contemporary mathematics, viscosity is completely lost in the prevailing Dedekind-Cantor theories of real analysis, and continuity generally. A line just is a set of points, and with the background classical logic (and set theory), every non-empty subset of a given line has a non-empty complement. One might say that the Dedekind-Cantor line is fragile, easily broken at every point, with nothing lost. Of course, the intermediate value theorem holds.

On the other hand, viscosity, or at least indecomposibility, is maintained in intuitionistic analysis and smooth infinitesimal analysis, but the intermediate value theorem fails there. The main theme of smooth infinitesimal analysis is that all functions are smooth: they are differentiable, their derivatives are differentiable, etc. Arguably, this is a natural extension (or consequence) of viscosity.

John Bell (2008) argues that this mantra—that all functions are smooth—is incompatible with the intermediate value theorem. He gives an example of a cubic polynomial, with two parameters ($x^3 + bx + c$). For each instance of the parameters b and c, there will be some positive and some negative values of the polynomial. So, if the intermediate value theorem held, there would be a projection function that takes the parameters as arguments and returns a zero of the polynomial with those parameters. Then Bell points out that no projection function of this polynomial is smooth.

So, we have to choose between two rather intuitive and longstanding principles once thought to underlie continuity. The natural extension of viscosity vs. the intermediate value theorem—just as we must choose between the Part-Whole Principle and Hume's Principle. Those who

adopt the prevailing Dedekind-Cantor account of continuity make one choice, those who favor intuitionistic analysis and smooth infinitesimal analysis make the other.

5. Collections, sets, classes, totalities, . . .

Arguably, the word "set" is now a technical term in mathematics and philosophy, referring to the iterative hierarchy, as encapsulated by Zermelo Fraenkel set theory with choice (ZFC). Perhaps "class" is also a technical term, as in "proper class". So let us settle on the more common term *collection*. Is that notion, or was that notion, subject to open texture, or was it governed by intuitive but mutually inconsistent principles? Or is the notion of collection monolithic, completely determined? Does it correspond, exactly, to the notion of "set" codified in ZFC? George Boolos (1988) seemed to think so. In a witty attack on a proposal to admit so-called "proper classes" to the theory, he wrote:

> Wait a minute! I thought that set theory was supposed to be a theory about all, "absolutely" all, the collections that there were and that "set" was synonymous with "collection". If one admits that there are proper classes at all, oughtn't one to take seriously the possibility of an iteratively generated hierarchy of collection-theoretic universes in which the sets which ZF recognizes play the role of ground-floor objects? I can't believe that any such view of the nature of [membership] can possibly be correct. Are the reasons for which one believes in classes really strong enough to make one believe in the possibility of such a hierarchy?

Sets, as codified in ZFC, are, of course, well-founded. Are there non-well-founded collections, presumably collections that are not sets? In particular, can a collection be a member of itself? Perhaps. Suppose that I am inspired by Julie Andrews's performance in *The sound of music* and elaborate a collection of my favorite things. Like Julie Andrews's character, the items on the list comfort me when I'm feeling sad.

The only item on my list that is also on Julie Andrews's list is warm woolen mittens, a must for anyone who lives in certain climates (such as Ohio and Connecticut). My list includes my iPad, my house, my atomic watch, my wife's record collection, my wife herself, my children and grandchildren, some of my colleagues, and my old running shoes—they are still very comfortable. There are also some abstract objects among my favorite things. They include the number 17 (because it is prime), the number 34 (because it is composite), the number $33,550,336$ (because, apparently, it is the smallest perfect number not known by ancient mathematicians), the pair $\langle 101, 103\rangle$ (a pair of twin primes), democracy, and free speech.

As it happened, I grew fond of the collection of my favorite things. I found that thinking about this collection (as opposed to its members) also sometimes comforts me when I am feeling sad. So I add the collection of my favorite things to the collection of my favorite things.

I presume that the gentle reader saw that coming. There does not appear to be a contradiction in my thinking here, at least not yet. Unless someone insists that comprehension or separation is somehow a constitutive principle of the very notion of "collection", but why think that? To channel Waismann, most notions are not articulated in every conceivable direction.

In one of his celebrated philosophical papers, "What is Cantor's continuum problem?", Kurt Gödel (1983) writes that there are two distinct intuitive or pre-theoretic notions of "set" or, in present terms "collection". One is when a collection is somehow tied to a property. A set, in this

sense, is something like a property in extension, perhaps what Frege called a "concept". This seems to be the notion behind Frege's "course of values" or "extension" operator, which leads naturally (but tragically) to contradiction via Basic Law V.

According to Gödel, the other notion of "collection" occurs when we start with a determinate totality of objects, such as the natural numbers, and consider collections of those—sets of numbers. And this notion iterates, we can talk about sets of sets of numbers, sets of sets of sets of numbers, etc. As Gödel put it in a footnote, the iteration continues into the transfinite. He concludes:

> As far as sets occur in mathematics ... they are sets of integers, or of rational numbers ... or of real numbers ... or of functions of real numbers ... This concept of set, however, according to which a set is something obtainable from the integers (or some other well-defined objects) by iterated application of the operation "set of", not something obtained by dividing the totality of all existing things into two categories, has never led to any antinomy whatsoever; that is, the perfectly "naïve" and uncritical working with this concept of set has so far proved completely self-consistent. (Gödel, 1983, §2, 474-5)

This is sometimes called the *iterative* notion of set, explicitly articulated in Zermelo (1930) and, of course, others. It is indeed natural to conclude that iterative sets are well-founded, and thus cannot be members of themselves. As Gödel puts it, an iterative set is "formed" from some pre-existing or otherwise available things. So it can't contain itself as a member. It follows that the collection of my favorite things is not an iterative set, at least not when I put that very collection into that very collection. But does that disqualify it from being a collection of a different sort?

In some moods, I am tempted to include the entire iterative hierarchy, V, in the collection of my favorite things, although I must admit that thinking about V does not always comfort me when I am feeling sad. Sometimes it just confuses me. Besides, it is also not so clear that V is a thing, and so it may not be eligible to be a favorite thing. If V is a thing, surely it is a collection, and not an iterative one.

Our question now is whether Gödel and Boolos are right that a monolithic, fully consistent, and completely determinate conception of collection was in place in mathematics all along, or whether the notion of collection is more like polyhedron, area, computability, and same-size, at least as characterized above.

In other words, is the iterative notion of set enshrined in ZFC a sharpening of a prior notion that had some open texture or perhaps was governed by inconsistent principles? What are we to make of the vigorous mathematical work on other kinds of set theories? There is work on non-well-founded sets (Aczel, 1988), and set theories that have a universal set, such as Quine's (1937) "New foundations" (see Forster, 1995), not to mention work on inconsistent set theory (Weber, 2009).

Is this apparently mathematical work simply a mistake, putting forward principles that are false of the given notion? Or are these other, perhaps less successful, accounts different ways to sharpen a pre-theoretic notion that was subject to open texture?

6. Upshot: in lieu of a conclusion

Recall the passage from Waismann (1951, 53):

> [T]here are no precise rules governing the use of words like 'time', 'pain', etc., and ... consequently to speak of the 'meaning' of a word, and to ask whether it has, or has not changed in meaning, is to operate with too blurred an expression.

The "too blurred expression" here is something like "has the same meaning as", or, in our cases, "is the same notion as" applied across contexts separated by significant temporal intervals. The thesis here is that the question as to whether the notions from sufficiently different times are the same or different is not always a good question. Without more context filled in, there might not be a determinate answer.

I submit that open texture, once it is discovered, and contradictory principles, once those are discovered, are resolved through semantic *choice*. The relevant linguistic/scientific/mathematical community collectively *decides* how to go on. The Waismann-inspired thesis is that after such choices are made, it is not always determinate whether the result is a change in the meaning of a given term or concept (*pace* one of the main arguments of LaPorte (2004)). To put the point in more Wittgensteinian terms, it is not always determinate whether the relevant intellectual community goes on as before after open texture is at least partially resolved.

Our question here is how do we make these kinds of choices? And who are "we"? We must get speculative. The proposal to be explored here is that it is a matter of what is now called conceptual engineering, or something similar like providing a Quine/Carnap explication (if either of those is different from conceptual engineering). Not to undermine or simply ignore a growing literature on this, what are the rules of that enterprise, in the case of mathematical notions that are up for sharpening or explication or ...? Our primary example, of course, is the notion of collection, as presented in the previous section.

Consider a remark that Michael Dummett (1991, 316) once made about the notion of "cardinal number", once we ponder the transfinite:

> We can gain some grasp on the idea of a totality too big to be counted ...but once we have accepted that totalities too big to be counted may yet have numbers, the idea of one too big even to have a number conveys nothing at all. And merely to say, "If you persist in talking about the number of all cardinal numbers, you will run into contradiction", is to wield the big stick, but not to offer an explanation.

Presumably, the explanation here would somehow invoke the pre-theoretic notion of "cardinal number", as it applies to infinite collections. However, a main theme here is that in cases like this, it is sometimes indeterminate just how the pre-theoretic notion applies to the new cases. Again, can we appeal to the Euclidean Part-whole principle, or do we favor Hume's Principle, or perhaps something else entirely instead?

Our current question is this: when trying to develop rigorous theories of notions subject to open texture and/or other sorts of indeterminacy—when we are trying to sharpen words or concepts—when is it acceptable to simply wield the Big Stick (even if one speaks softly) and when is a more satisfying explanation required (or at least strongly desired)? If an explanation is required, it is presumably to come from the pre-theoretic concept.

Let P be a target notion. There a number of possibilities. Here is one:

> **Scenario 1**: P is monolithic, in the sense that it has no open texture, is completely coherent, and is not up for sharpening, at least not in different ways.

This seems to be Gödel's and Boolos's view of the present case. They claim (or presuppose) that there always was a single, coherent notion of collection, or at least of (iterative) set, and we are invoking that very notion today.

In a case like Scenario 1, we are up for traditional conceptual analysis, not conceptual engineering, explication, or the like. It is a matter of discovering just what the monolithic phrase or concept *is*. Wielding the Big Stick is clearly inappropriate. We are looking for an explanation, not an explication and thus not a replacement or improvement or sharpening. In the present case, for example, we might be looking to explain why all sets are well-founded and thus why there is no universal set. All proposed axioms and principles should be justified by the nature of P, whatever that nature should turn out to be.[7]

On to a second possibility:

> **Scenario 2:** P is not monolithic—it has some open texture or is governed by mutually inconsistent principles, or something similar. But it turns out that there are two (or three or four) notions that somehow underlie it, and each of those underlying notions is itself monolithic: each has no open texture and each is completely coherent.

As we have seen, when it comes to "collection" Gödel did not say that we are (or were) in this Scenario with respect to the notion of "collection". He did say that there are two notions, but he insisted that one of them—the one based on properties—is incoherent, and so of no mathematical interest. He directed our attention to the other notion of collection, the iterative one, which he did regard as monolithic. Gödel did famously say that once we realize the our target notion is that of iterative set, the axioms of ZFC "force themselves on us" as true. Moreover, he claimed that the iterative notion is the the *only* notion of collection in use in mathematics.

Luca Incurvati (2021) articulates an interesting and insightful program much like Scenario 2 concerning our target notion of set (or collection). He begins (pp. 16-17) by laying out a (or the) pre-theoretic, or intuitive, "concept" of set which, he explicitly notes, is subject to open texture, citing Waismann (and Shapiro, 2006a). A "conception" of set is a further articulation of the underlying concept, filling in resolutions to open texture (and other indeterminacies) in various ways. A "conception" of set is at least of-a-piece with what we here call a "sharpening" of the notion (or concept).

Incurvati develops several "conceptions" of set in delightful metaphysical, epistemic, and mathematical detail, including an iterative conception, a dialetheic naïve conception invoking unrestricted comprehension and a paraconsistent logic, a "limitation of size" conception, a stratified conception based on W. V. O. Quine's (1937) "New foundations for mathematical logic", and a "graph" conception based on Aczel's (1988) non-well-founded set theory. After laying out criteria for choosing among conceptions, along with the strengths and weaknesses of each of the conceptions, Incurvati settles on the iterative conception via an "inference to the best conception". Of course, one might ask, best for what?

In linguistic terms, one way to describe a case like that of Scenario 2 is to say that that the word or phrase for P is polysemous: it has more than one meaning, although the meanings are

[7] We are not concerned here with specific statements about sets, such as the continuum hypothesis, whose truth value is not determined by what we have assumed about the (alleged) monolithic notion. Clearly these matters are related.

related to each other in systematic ways. If each of the sharpenings (or meanings) is itself free of open-texture and the like, then we have two (or three, four, or, in Incurvati's case, five) instances of Scenario 1.

Like Scenario 1, there is no conceptual engineering or explication in Scenario 2. *Each* of the underlying notions is up for traditional conceptual analysis, and so there should be no wielding of the Big Stick. Any proposed axioms or principles for any of the underlying notions (or, in Incurvati's terms, conceptions) should be explained by that underlying notion, assumed to be coherent and sufficiently determinate.

And now for our third possibility:

> **Scenario 3:** P is not monolithic—it has some open texture or is governed by mutually inconsistent principles, or something similar. But in this case, there is no clear candidate for an intuitive, pre-theoretic and coherent notion or notions, that somehow underlie P. The notion P can be sharpened in various ways that are incompatible with each other. And these underlying notions are still subject to open texture (perhaps in different ways).

Here P *is* up for conceptual engineering, explication and/or the like, possibly in more than one way. In this kind of case, we have to look at the *purposes* for which the conceptual engineering and/or explication is done. After all, with any engineering project—conceptual or otherwise—one proposes to make (or articulate) something that is to serve a particular purpose. We make this in order to accomplish that. Well, just what is the purpose (or purposes) of set theory, our mathematical theory of collection?

Similarly (or identically), a Carnap/Quine explication is a proposal that a certain articulated notion is to replace an intuitive or pre-theoretic one *for certain purposes*. So the explicator should say what the purposes are, and then show how those purposes can be used to guide what the replacement, or replacements, might be. They should show just how the replacement accomplishes the stated purpose.

Let us assume that something resembling Scenario 3 makes sense (or did make sense in the past) with respect to our candidate notion of collection. So we are (or were) up for conceptual engineering or explication. For what purpose(s)? What is the collection-theory to do? What is it for?

These, of course, are interesting and perhaps complicated questions, calling for speculation on what the purposes of various mathematical theories are, presumably assuming that each theory has a single purpose. As a first answer, one *might* just think of set theory as like any other mathematical theory. Then we should perhaps adopt the now common Hilbertian theme that consistency is the only legitimate criterion that a mathematical theory must meet.[8]

The usual model here is geometry. As Alberto Coffa (1986, 8, 17) once put it:

> During the second half of the nineteenth century, through a process still awaiting explanation, the community of geometers reached the conclusion that all geometries were here to stay … [T]his had all the appearance of being the first time that a community of scientists had agreed to accept in a not-merely-provisory way all the members of a set of mutually inconsistent theories about a certain domain.

[8] If dialetheism is not to be set aside, then perhaps non-triviality is the relevant criterion.

The idea is that, as mathematics, all of the geometries are legitimate theories. There is no more conflict between Euclidean, spherical and hyperbolic geometries than there is between arithmetic and real analysis. They are just different mathematical theories—about different kinds of things.

Of course, there is still a question of whether a given theory is interesting, or theoretically fruitful. But perhaps we can just continue this quasi neo-pragmatist query: fruitful for what purpose? There is also an interesting and important question of which mathematical theory is best employed in a given scientific theory. But the question of which geometry to employ in physics is not a mathematical question, or at least it is not a question for mathematicians, qua mathematicians, to settle. It is a matter of applied mathematics. When it comes to so-called pure mathematics, Coffa is correct: all of the geometries are here to stay.

So *if* a given collection theory is to be understood as like any mathematical theory, then mathematicians can simply choose whatever axioms they want to explore. There can be, and indeed, there are, non-well-founded theories, iterative theories, theories with and without various axioms of choice or various determinacy principles, theories that accept the continuum hypothesis, or V=L, or various large cardinals, not to mention theories using the Part-Whole principle (for the associated notion of cardinality), etc.

If, following Hilbert, consistency is the criterion—indeed the only criterion— for a theory of collection, then it *is* appropriate to simply wield the Big Stick. From this perspective, if one is asked to explain why it is that there is no universal set in ZFC, the only answer is that this follows from the axiom of foundation or, if you like, the axiom of separation. One simply declares that this is the theory I am interested in developing (putting applications aside, as in much of mathematics). If seemingly contrary decisions could have been made, one is free to make them instead, just as in geometry (or at least the emerging view of geometry toward the end of the nineteenth century). As we have seen in the explosion of work in set theory (or set theories), different decisions were made and have been (and continue to be) explored in detail.

Hilbert did declare that intuition is the source of the axioms of (Euclidean) geometry, and it is undeniable that at least some of the axioms of ZFC (or the other set theories) do conform to intuitive or pre-theoretic concepts. The key observation in both cases is that once the axioms are chosen, they alone guide the mathematical development of the theory.

It might be added that in some cases, applications for a given theory were found only after the theory had been developed and explored by mathematicians. In general, one cannot always tell, in advance, whether a given theory will prove fruitful for this or that extra-mathematical purpose (see Steiner, 1997), or whether it will continue to hold of some intuitive or pre-theoretic concept or idea, especially if the pre-theoretic concept or idea is fraught with open texture. So perhaps the best policy is to let a thousand flowers bloom, or at least to let a thousand flowers try to bloom, even if not many do.

On the other hand (and finally), it is sometimes thought that set theory is *not* just another mathematical theory like any other. It has a certain *foundational* role in mathematics. It is, of course, controversial just what this foundational role is (and also whether mathematics needs a foundation). This is not the place to engage that matter in any detail, but we can sketch it (see Shapiro, 2004).

For some, a foundation is metaphysical. One might think that set theory provides the ultimate ontology of all of mathematics: natural numbers, real numbers, geometric spaces, functions, graphs, and the like really are sets (ultimately). In another sense, perhaps, a foundation is

epistemological—it tells what mathematical knowledge (in, say, analytic function theory) is—it is really or ultimately knowledge about the universe (or universes) of sets.

Neither of these senses is all that relevant here. The foundation we have in mind is more internal to mathematics. Here is Penelope Maddy's (2007, 354) summary of some of the relevant foundational features of set theory:

> … set theory hopes to provide a dependable and perspicuous mathematical theory that is ample enough to include (surrogates for) all the objects of classical mathematics and strong enough to imply all the classical theorems about them. In this way, set theory aims to provide a court of final appeal for claims of existence and proof in classical mathematics … Thus set theory aims to provide a single arena in which the objects of classical mathematics are all included, where they can be compared side-by-side. Given this foundational goal, … set-theoretic practice must strive to settle on one official theory of sets, a single fundamental theory. This is not to say that alternative set theories could not or should not be studied, but their models would be viewed as residing in one true universe of sets, V.

In a case like this, where the conceptual engineering or explication is to further a specific goal—in this case a foundational one in the relevant sense—it seems that it is not sufficient to simply wield the Big Stick when choosing axioms or constitutive principles. Clearly, however, we are not looking for a traditional conceptual analysis either, not in terms of an intuitive or pre-theoretic concept (or a conception in Incurvati's sense). The details of the theory need not answer to a pre-existing monolithic notion of collection (or anything else). Instead, we are looking for an explanation of the proposed axioms in terms of the foundational goals of the theory.

Indeed, following Maddy, any decisions to be made in generating the theory should be addressed by how best to further the stated foundational goal of the enterprise. This, of course, is just what Maddy and others do concerning matters like V=L and some large cardinal hypotheses.[9] I have no insights to offer on what the rules of this enterprise are.

I conclude with an observation that, in this context, we have some reason to prefer a potentialist version of set theory, along the lines of Linnebo (2013), Hellman (1989), or others.

Of course, ZFC has no universal set—V is not a set. Again, if we think of ZFC as a mathematical theory like any other, then, as above, there is no need to explain this. The lack of universal set is a consequence of the axioms that characterize the theory.

But if set theory is to serve its foundational role, then perhaps we do need an explanation. It is not enough to wield the Big Stick. As noted, Gödel and others explain the lack of a universal set in terms of the iterative notion of set. Each set is "formed" from previously available objects (typically other sets). There is no stage when all sets are "available". Each stage can be surpassed.

These proposed explanations are thus tied to the presumably pre-theoretic notion of iterative set, and do not address the foundational role of the theory, at least not directly.

As Maddy notes, the foundational theory should have isomorphic surrogates for *every* legitimate mathematical structure. And if set theory is to be a legitimate mathematical theory, it

[9] It is thus quite relevant that Maddy explicitly prefers "extrinsic" justifications for axioms, based on the foundational goals, to "intrinsic" justifications which relate to the underlying concept or, in Incurvati's terms, conception of set.

should have a surrogate for its own domain, V. Surrogates are always sets, so V should be a set (or at least be represented by one).

Perhaps Maddy's talk of models of set theory (which also presumably reside in V) can play this role, but, as is well-known, set theory cannot prove that it has a model (in light of the second incompleteness theorem). And it can be proved that no set has, as members, all of V. Thus, there is no surrogate for V itself, at least if we assume that surrogates are supposed to isomorphic copies.

One option, perhaps, would be to back off slightly from the foundational claim, and insist that set theory serves as a foundation for all of mathematics except set theory itself (and perhaps other proposed foundations, like the category of all categories). If we accept the prevailing foundational role, then set theory remains different from every other mathematical theory. If nothing else, there is no surrogate—no isomorphic copy—of its own domain in the iterative hierarchy (thanks to the axiom of foundation and/or the axiom of separation).

A better option, I propose, it to adopt a potentialist set theory for this foundational purpose, perhaps in a background modal language (following Hellman, 1989; Linnebo, 2013, or others). So what is the explanation of why there is no universal set according to this theory? We do not invoke a prior, intuitive or pre-theoretic concept or conception of collection. Rather, we argue that there is no universal set because mathematics does not and cannot have a single domain for all of its present and future endeavors.

In other words, there is a kind of indefinite extensibility to mathematics itself—there simply is no fixed domain for all of it. Clearly, one might postulate a single domain—a single set—that has surrogates for all of the mathematics *at a given time*. If we focus on the period just before ZFC came along, for example, surely $V_{\omega+4}$, or so, would do (depending on how much coding one wants to include). But as soon as this particular domain is countenanced, mathematics will go beyond it, and encounter richer domains. If nothing else, we have its powerset: $V_{\omega+5}$.

It is clear that mathematics has gone well beyond Aristotle's rejection of actual infinity. Apparently, mathematics has also gone beyond Aristotle's teacher and main opponent. Plato was critical of the geometers of his day, arguing that their dynamic, or constructive language is inconsistent with the nature of the true subject matter of geometry:

> [The] science [of geometry] is in direct contradiction with the language employed by its adepts ... Their language is most ludicrous, ... for they speak as if they were doing something and as if all their words were directed toward action ... [They talk] of squaring and applying and adding and the like ... whereas in fact the real object of the entire subject is ... knowledge ... of what eternally exists, not of anything that comes to be this or that at some time and ceases to be. (*Republic*, VII)

We can leave it to others to debate whether this Platonic thought accommodates the mathematics of Plato's day, or of any given time slice of mathematics. But it is getting clear that there can be no timeless, static, and eternal World of Being that somehow contains all that there is or ever will be in mathematics.

References

Aczel, P. (1988). *Non-well-founded sets*. CSLI Lecture Notes: Number 14. CSLI Publications, Stanford.

Bell, J. (2008). *A primer of infinitesimal analysis*. Cambridge University Press, Cambridge, 2 edition. DOI: https://doi.org/10.1017/CBO9780511619625.

Blackburn, S. (1994). *The Oxford dictionary of philosophy*. Oxford University Press, Oxford.

Boolos, G. (1988). Reply to Charles Parsons' "Sets and classes". In *Logic, logic, and logic*, pages 30–36. Harvard University Press, Cambridge, Massachusetts.

Coffa, A. (1986). From Geometry to tolerance: sources of conventionalism in nineteenth-century geometry. In *From quarks to quasars: philosophical problems of modern physics*, volume 7 of *University of Pittsburgh Series*, pages 3–70. Pittsburgh University Press, Pittsburgh.

Dummett, M. (1991). *Frege: Philosophy of mathematics*. Harvard University Press, Cambridge, Massachusetts.

Forster, T. E. (1995). *Set theory with a universal set: exploring an untyped universe*. Oxford University Press, Oxford, 2 edition. DOI: https://doi.org/10.1093/oso/9780198514770.001.0001.

Gödel, K. (1983). What is Cantor's continuum problem? In Benacerraf, P. and Putnam, H., editors, *Philosophy of mathematics*, pages 470–485. Cambridge University Press, Cambridge, 2 edition.

Hellman, G. (1989). *Mathematics without Numbers: Towards a Modal-Structural Interpretation*. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/0198240341.001.0001.

Hellman, G. and Shapiro, S., editors (2021). *The history of continua*. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/oso/9780198809647.001.0001.

Incurvati, L. (2021). *Conceptions of set and the foundations of mathematics*. Cambridge University Press, Cambridge. DOI: https://doi.org/10.1017/9781108596961.

Lakatos, I. (1976). *Proofs and refutations*. Cambridge University Press, Cambridge.

LaPorte, J. (2004). *Natural kinds and conceptual change*. Cambridge University Press, New York. DOI: https://doi.org/10.1017/CBO9780511527319.

Linnebo, Ø. (2013). The potential hierarchy of sets. *Review of Symbolic Logic*, 6:205–228. DOI: https://doi.org/10.1017/S1755020313000014.

Maddy, P. (2007). *Second philosophy: a naturalistic method*. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/acprof:oso/9780199273669.001.0001.

Mancosu, P. (2009). Measuring the size of infinite collections of natural numbers: was Cantor's theory of infinite number inevitable? *Review of Symbolic Logic*, 2:612–646. DOI: https://doi.org/10.1017/S1755020309990128.

Mendelson, E. (1963). On some recent criticisms of Church's thesis. *Notre Dame Journal of Formal Logic*, 4:201–205.

Oliver, A. and Smiley, T. (2013). *Plural logic*. Oxford University Press, Oxford, 2 edition. DOI: https://doi.org/10.1093/acprof:oso/9780198744382.001.0001.

Plato (1963). *The collected dialogues*. Princeton University Press, Princeton.

Quine, W. V. O. (1937). New foundations for mathematical logic. *American Mathematical Monthly*, 44:70–80.

Shapiro, S. (2004). Foundations of mathematics: metaphysics, epistemology, structure. *Philosophical Quarterly*, 54:16–37.

Shapiro, S. (2006a). Computability, proof, and open-texture. In Olszewski, A., Woleński, J., and Janusz, R., editors, *Church's thesis after 70 years*, pages 420–455. Ontos Verlag, Frankfurt. DOI: https://doi.org/10.1515/9783110325461.420.

Shapiro, S. (2006b). *Vagueness in context*. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/acprof:oso/9780199280391.001.0001.

Shapiro, S. and Roberts, C. (2021). Open texture and mathematics. *Notre Dame Journal of Formal Logic*, 62:173–191. DOI: https://doi.org/10.1215/00294527-2021-0007.

Smith, S. R. (2015). Understanding of concepts: the case of the derivative. *Mind*, 124:1163–1199. DOI: https://doi.org/10.1093/mind/fzv068.

Steiner, M. (1997). *The applicability of mathematics as a philosophical problem*. Harvard University Press, Cambridge, Massachusetts.

Waismann, F. (1945). Verifiability. *Proceedings of the Aristotelian Society*, Supplementary Volume 19:119–150. Reprinted in *Logic and Language*, edited by Antony Flew, Oxford, Basil Blackwell, 1968, pp.117-144.

Waismann, F. (1951). Analytic-synthetic iii. *Analysis*, 11:49–61.

Weber, Z. (2009). Inconsistent mathematics. http://www.iep.utm.edu/math-inc/. Internet Encyclopedia of Philosophy.

Youschkevitch, A. J. (1976). The concept of function up to the middle of the nineteenth century. *Archive for History of Exact Sciences*, 16:37–85.

Zermelo, E. (1930). Über Grenzzahlen und Mengenbereiche: Neue Untersuchungen über die Grundlagen der Mengenlehre. *Fundamenta Mathematicae*, 16:29–47. Translated as "On boundary numbers and domains of sets: new investigations in the foundations of set theory", in *From Kant to Hilbert: a source book in the foundations of mathematics, Volume 2*, edited by William Ewald, Oxford, Oxford University Press, 1996, 1219-1233.

Journal for the Philosophy of Mathematics

Firenze University Press
www.fupress.com

The Plural Iterative Conception of Set

DAVIDE SUTTO

University of Oslo, Faculty of Humanities, Department of Philosophy (IFIKK)
Email: davide.sutto@ifikk.uio.no

OPEN ACCESS

Citation: SUTTO, Davide (2025). The Plural Iterative Conception of Set. *Journal for the Philosophy of Mathematics*. 2: 161-193. DOI: 10.36253/jpm-3549

Received: June 14, 2025

Accepted: October 06, 2025

Published: December 30, 2025

ORCiD
DS: 0000-0003-4138-8438

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Abstract: Georg Cantor informally distinguished between "consistent" and "inconsistent" multiplicities as those many things that, respectively, can and cannot be thought of as one, i.e., as a set. To clarify this distinction, the recent debate filtered the logic of plurals through two main approaches to the process of set-formation: limitation of size (Burgess) or set-theoretic potentialism (Linnebo). In this paper I propose a third route through the development of a plural iterative conception of set. Inspired by Tim Button's Level Theory, I define and axiomatize the notion of a plural level, which explains Cantor's multiplicities either as level-bound (consistent) or level unbound (inconsistent) pluralities. While this framework is clearly in contrast with the limitation of size view, it also revives a plausible actualist picture prematurely dismissed by the advocates of potentialism.

Keywords: Iterative conception, set theory, plural logic, Cantor, level theory.

1. Introduction

In 1895 Georg Cantor opened one of his articles with the following definition of set:

> By an "aggregate" [*Menge*] we are to understand any collection into a whole [*Zusammenfassung zu einem Ganzen*] M of definite and separate objects m of our intuition or our thought. These objects are called the "elements" of M. In signs we express this thus: $M = \{m\}$. (Cantor, 1895, p. 85)

The reading of 'm' as a plural variable, 'mm' in modern notation, seems quite natural and has already been extensively defended in the literature.[1]

The idea that sets are obtained by abstraction from a plurality of definite and fixed objects had already been pursued by Cantor in the *Grundlagen*:[2]

[1] See Burgess (2004); Florio and Linnebo (2021); Linnebo (2010, 2013); Oliver and Smiley (2016, 2018).

[2] One could also interpret this definition as opposed to the previous one, in that Cantor seems to be moving from a logical (1883) to a combinatorial (1895) notion of set. In fact, "united into a whole by some law" seems to suggest a conception where sets are pinned down by conditions 'ϕ' and not just by the plurality of their elements, as it is instead clear from the above definition. However, one could still take "aggregate of determinate elements" to be something like a proto-notion of a plurality. This is even more so in light of the interpretation advanced in the present paper, which endorses the full principle of comprehension for pluralities, bringing them closer to proper classes.

> In general, by a 'manifold' or 'set' I understand every *multiplicity* [jedes Viele] *which can be thought of as one*, i.e. every aggregate [Inbegriff] of *determinate elements* which can be *united* [verbunden] *into a whole by some law*. (Cantor, 1883, p. 916, my emphasis)

The same thought resurfaces also in two famous letters to Hilbert and Dedekind dated, respectively, 1897 and 1899:

> I say of a set that it can be thought of as *finished* [...] if it is possible without contradiction (as can be done with finite sets) to think of *all its elements as existing together* [...]; or (in other words) if it is possible to imagine *the set as actually existing with the totality of its elements*. [...] And so too, in the first article of [Cantor (1895)], I define a 'set' [...] at the very beginning as an 'assembling together' [Zusammenfassung]. But an *'assembling together'* is only possible if an *'existing together'* [Zusammensein] is possible. (Cantor, 1897, pp. 927–928, my emphasis)

> If we start from the notion of a *definite multiplicity* [Vielheit], [...] it is necessary [...] to distinguish two kinds of multiplicities (by this I always mean *definite* multiplicities). For a multiplicity can be such that the assumption that *all* of its elements 'are together' leads to a contradiction, so that it is impossible to conceive of the multiplicity as a unity, as 'one finished thing'. Such multiplicities I call *absolutely infinite or inconsistent multiplicities*. [...] If on the other hand the *totality of the elements* of a multiplicity can be thought of without contradiction as 'being together', so that they can be gathered together into '*one thing*', I call it a *consistent multiplicity* or a 'set'. (Cantor, 1899, pp. 931–932, my emphasis)

We can extract two main ideas from these passages. The first is that Cantor conceived *set-formation* as *abstraction from pluralities (i.e., multiplicity, Zusammensein) to sets (i.e., one thing, Zusammenfassung)*:[3] $mm \mapsto \{mm\}$. The second is that not any "*multiplicity*" can form a set, but only those that are "*consistent*". Cantor spells out this notion in various ways and some become more perspicuous if read through the contemporary development of plural logic after George Boolos' seminal work.

Plural logic has been traditionally developed as a form of higher-order (read "second-order") quantification, hence tied to a Principle of Comprehension, which generates pluralities from any given condition ϕ (with xx not free):

$$\exists xx \forall x (x \prec xx \leftrightarrow \phi(x)) \qquad \text{(P-COMP)}$$

As Linnebo (2010); Yablo (2006) clearly show, a version of Russell's Paradox quickly emerges if we take only Cantor's first idea without the second, namely if we don't restrict our pluralities to "consistent multiplicities", whatever it may mean.[4] That is, if we assume that any plurality whatsoever forms a set, formally

[3] This could also be framed in terms of the conception of "*set-as-one*", to be contrasted with the "*set-as-many*" conception defended by Stanisław Leśniewski in the early days of the discipline (see Potter, 1990, §1.1 and Fraenkel, Bar-Hillel, and Lévy, 1973, §11.1) and clearly articulated already by Russell in *The Principles of Mathematics*, §74 (1903. See also Klement, 2014).

[4] Cantor was well aware of this, although he focused on the plurality of all cardinal numbers, which, if collapsed, leads to the paradox that bears his name:

$$\forall xx \exists x(x = \{xx\}) \qquad \text{(COLLAPSE)}$$

it suffices to plug in the plurality 'rr' of all non-self-membered sets, generated by P-COMP, to obtain the antinomic Russellian set.[5] Therefore, coming back to his explanation of "(in-)consistent multiplicity", to formulate a coherent account of sets *qua* obtained from pluralities we need to make Cantor's original idea more precise. While the recent literature clarified these notions either in terms of a limitation of size view (Burgess, 2004; Pollard, 1996) or by outlining a potentialist account of sets (Florio and Linnebo, 2021; Linnebo, 2013), it seems that a gap has been left open: what about the (non modal) Iterative Conception of Set?[6]

The present paper fills this gap and explores a way of clarifying Cantor's distinction through the means offered by the most popular conception of set (Incurvati, 2020). Despite its great popularity, the vast majority of axiomatizations of the Iterative Conception are first-order and overlook the first Cantorian idea of conceiving the process of set construction as a process of plural-to-set abstraction. My aim is to axiomatize a conception of set that sharpens Cantor's ideas through means (the iterative picture) that are not to be attributed directly to him (see Ferreirós, 2007, Epilogue) and that, in the end, sanction the usual axioms of Zermelo-Fraenkel set theory. As Oliver and Smiley observe, "historians agree that an iterative conception of sets is, as doctors say in court, consistent with Cantor's ideas" (Oliver and Smiley, 2016, p. 265), which legitimizes the project of combining the Cantorian view of sets with the Iterative Conception. While Oliver and Smiley (2016, 2018) also advance a theory that goes in this same direction,[7] their aim is to do justice to what they consider a full-fledged "Cantorian orthodoxy" that rejects things such as singleton and empty sets to match their account of pluralities. Therefore, despite a similar appeal to an axiomatization of the Iterative Conception in terms of levels, our respective projects are different and shall not be confused.[8]

The paper is structured as follows. In §2 I introduce the language of plurals and show how it can articulate the process of set formation. In §3 I introduce the raw idea of the Iterative Conception of Set and explain how it can be integrated with the idea that sets are obtained from pluralities. §4 and §5 are the core of the paper, where I propose an axiomatization of the conception under the label of Plural Level Theory (PLT) and where I introduce its most salient results. In §6 I compare PLT with alternative approaches to the development of a Cantorian view of sets. §7 closes the paper with a summary of the main results. In Appendix A I gather the most technical details of the framework.

2. From Pluralities to Sets

Starting from the overall idea of grounding set theory into plural talk, I proceed to define a plural language and see how it interacts with sets.

> In contrast, infinite sets such that the *totality* of their elements cannot be thought of as 'existing together' [...], and that therefore also *in this totality* are absolutely not an object of further *mathematical* contemplation, I call *'absolutely infinite sets'* and to them belongs the 'set of all alephs'. (Cantor, 1897, pp. 927–928, his emphasis)

[5] To be absolutely precise one would also need a principle of set-abstraction that grants that identical sets are mapped to identical pluralities, formally '$\{xx\} = \{yy\} \leftrightarrow xx \approx yy$' (see below for the notation). In fact, in accounts like Florio and Linnebo (2021) this principle, being more fundamental, substitutes COLLAPSE from Linnebo (2010).

[6] Whether or not potentialism is an instance of the Iterative Conception is up for debate: Linnebo (2013); Studd (2013) answer positively, while Roberts (MS) frames it as an alternative both to the conception and to Limitation of Size.

[7] Another theory combining the Iterative Conception with plural quantification is presented by Pollard (1985). However, his work is radically different both from mine and from Oliver and Smiley's since he starts from the ordinals as urelements and investigates how second-order quantification *à la* Boolos can make sense of various approaches to the conception.

[8] If one wants to proceed with a comparison, in §6.3. I argue that my approach fares better also if framed within their "Cantorian orthodoxy".

2.1. *Pluralities*

$\mathcal{L}_{\prec}$ is an extension of first order logic (FOL) with identity that, along with plural variables '$xx, yy, zz, \ldots$' and quantifiers, introduces a plural "membership" predicate '$\prec$'. This takes singular terms on the left and plurals on the right, so '$x \prec xx$' is to be read as 'x is among/one of the xx'. $\mathcal{L}_{\prec}$ adds two axioms to FOL:

$$\exists xx \forall x(x \prec xx \leftrightarrow \phi(x)) \qquad \text{(P-COMP)}$$

$$\forall xx, yy(\forall u(u \prec xx \leftrightarrow u \prec yy) \rightarrow (\Phi(xx) \leftrightarrow \Phi(yy))) \qquad \text{(P-INDISC)}$$

The first is the principle P-COMP from above, which grants an unrestricted generation of "*definite*" pluralities, both "consistent" and "inconsistent". For me this is a crucial part of the Cantorian spirit of the project since Cantor seems to be explicitly committed to a free generation of "definite multiplicities", which I read as "*pluralities defined by a comprehension principle*". The crucial point is then demarcate between those definite multiplicities that can and those that cannot collapse into sets. Nonetheless, the acceptance of P-COMP, is a contentious point between my proposal and potentialist accounts of sets, both modal and non-modal (see §6.2.). The second axiom is just indiscernibility of identicals for pluralities.[9] My formulation follows Burgess (2004) and it is enough to preserve the extensional nature of pluralities in the sense that co-extensionality (left-hand-side) implies indiscernibility (right-hand-side).[10]

We start with the definition of the derived relation of (proper) *plural inclusion* (or *subplurality*) and *plural identity* as in Florio and Linnebo (2021, §2.3):[11]

Definition 2.1 (P-IND). [12]

(i) $bb \preccurlyeq aa \leftrightarrow_{def} (\forall x \prec bb)x \prec aa$
(ii) $bb \npreccurlyeq aa \leftrightarrow_{def} (\forall x \prec bb)x \prec aa \wedge \neg(aa \preccurlyeq bb)$
(iii) $bb \approx aa \leftrightarrow_{def} bb \preccurlyeq aa \wedge aa \preccurlyeq bb$

Just as I differentiate between plural and singular identity with '$\approx$' and '$=$', in the same way, I use '$:\approx$' in definitions of specific pluralities in place of '$:=$'.

[9] Florio and Linnebo (2021, p.19) use lowercase schematic letters specifying that '$\phi(xx)$' stands for the standard substitution of all free occurrences of some free variable vv by xx whenever vv can be substituted by xx in ϕ. Here I prefer to adopt capital letters to avoid the following misunderstanding, which one can easily imagine arising if one is not already familiar with the language of plurals: one may interpret $\phi(xx)$ as being the plurality of all x such that $\phi(x)$ characterized by P-COMP above. Due to pluralities being extensional entities (as can be easily derived from P-INDISC), there is just one such plurality (as in the case of sets) and, under this interpretation, it would make no sense to speak of "all pluralities that are $\phi(xx)$" as done in the axioms below. To avoid such misunderstanding, I adopt the present notation to tag schematically collective predicates.

[10] Alternatively, one could add a primitive plural-identity predicate '$\approx$' with the axioms: P-COMP +

$$\forall xx, yy(xx \approx yy \rightarrow (\Phi(xx) \leftrightarrow \Phi(yy))) \qquad \text{(P-INDISC}^{\approx}\text{)}$$

$$\forall aa \forall bb(\forall x(x \prec aa \leftrightarrow x \prec bb) \rightarrow aa \approx bb) \qquad \text{(P-EXT)}$$

This is similar to axiomatic set theory where Indiscernibility is a logical axiom and implies the right-to-left direction of Extensionality. Here we can dispense '$\approx$' as a further primitive and define it instead, while P-INDISC is already apt to do the job of Extensionality in granting that the notions defined below pin down pluralities in a unique way. Moreover, given Def. 2.1 below, we can easily derive P-INDISC$^{\approx}$. For this reason I go with the option more customary in the literature on plurals (Burgess, 2004; Florio and Linnebo, 2021; Oliver and Smiley, 2016). For more on plural primitives see Linnebo (2007, §3).

[11] They introduce these definitions before introducing P-INDISC, but the order should be reversed.

[12] For readability, I adopt the same conventions as Button (2021, §0). That is, I concatenate infix conjunctions and I abbreviate bounded quantifiers in the usual way. E.g., I write '$(\forall x \prec yy)\phi$' for '$\forall x(x \prec yy \rightarrow \phi)$' and '$(\forall x : \Psi)\phi$' for '$\forall x(\Psi(x) \rightarrow \phi)$'. The rest is the same for the existential quantifiers and any predicate 'Ψ' or infix predicates other than '$\prec$'.

2.2. Sets

Coming to the first Cantorian idea of conceiving sets as collapsed pluralities, I enrich $\mathcal{L}_{\prec}$ with an additional primitive non-logical symbol '$\ltimes$' of the reverse type of '$\prec$', where '$xx \ltimes x$' is to be read as "the plurality xx collapses into the set x". I denote the new language as $\mathcal{L}_{\prec,\ltimes} = \{\ltimes\}$. This allows set-membership to be defined:[13]

Definition 2.2 (P-IND). $x \in a \leftrightarrow_{def} (\exists aa \ltimes a)x \prec aa$, alternatively, $(\forall aa \ltimes a)x \prec aa$

To avoid making the notation too cumbersome from now on I will freely use '$\in$' and its derived relations '$\subseteq$ / $\subset$' thoroughly. However, bear in mind Def. 2.2 as making explicit Cantor's first idea concerning the process of set-formation.

The behavior of the collapse relation is further prescribed to be extensional:[14]

$$\forall aa, bb \forall a, b((aa \ltimes a \wedge bb \ltimes b) \rightarrow (a = b \leftrightarrow aa \approx bb)) \qquad (\text{EXT}_p)$$

An obvious consequence of EXT_p is that $(xx \ltimes a \wedge xx \ltimes b) \rightarrow a = b$, that is, if it exists, the collapse of a plurality into a set is uniquely determined (and vice-versa). This allows the introduction of functional notation '$\uparrow xx / \downarrow x$' to directly refer, respectively, to the collapse/uncollapse of a plurality/set as terms.[15] To differentiate between plural and set abstraction, I use the double pipe to tag the former, that is, I denote the set a of all the ϕ as '$a = \{x : \phi(x)\}$' and the respective plurality aa as '$aa \approx \|x : \phi(x)\|$'.[16] Combined together these yield other useful definitions:[17]

Definition 2.3 (P-INDISC, EXT_p).

(i) $\uparrow aa \coloneqq \{x : x \prec aa\}$
(ii) $\downarrow a :\approx \|x : x \in a\|$
(iii) $a \cap aa :\approx \|x : x \in a \wedge x \prec aa\|$
(iv) $a \sqsubseteq aa \leftrightarrow_{def} (\forall x \in a)x \prec aa$
(v) $a \sqsupseteq aa \leftrightarrow_{def} (\forall x \in aa)x \in a$
(vi) $a \blacktriangleright aa \leftrightarrow_{def} (\exists c \prec aa)a \subseteq c$

Concerning (1) and (2) note that, while $\downarrow a$ always exists if a does by Def. 2.2, the same is not true for $\uparrow aa$ and aa since this depends on the sharpening of "consistent multiplicity" we are after. In this sense, while '$\downarrow$' represents a total function, '$\uparrow$' stands for a partial function that sometimes is non-denoting. The aim of our theory can then be restated as singling out when '$\uparrow$' behaves like a total function. This is a further difference with system and Oliver and Smiley (2016), who simply accept the partiality of the 'set of' function (see §6.3. below).

Therefore, concerning Def. 2.3(1), whenever we have the collapse of a plurality into a set, but we miss the qualification of "consistency", we shall read that as: *if the collapsed set exists*, then it

[13] And also of a set-predicate, where needed: $ß(a) \leftrightarrow_{def} (\exists aa \ltimes a)$. This is the same as the axiom HEREDITY (3.1) from Burgess (2004, p. 198), who takes 'ß' as a primitive. I can define it because I am not appealing to a principle of reflection, as Burgess does, and so I am not worried about relativized formulae. Nonetheless, the definition of 'ß' has the same effect of HEREDITY, hence its name: all the (set)elements of a set are included in the universe of discourse as soon as the set is introduced

[14] This axiom is also present in Burgess (2004) as 3.2. However, since Burgess assumes a set-predicate as a further primitive (see fn. 13), he is forced to assume (or, better, derive) a principle of PURITY (7.2) to derive the usual set-theoretic axiom. On the contrary, here the set-theoretic axiom follows from P-INDISC, EXT_p and Def. 2.2, which already excludes ur-elements from the scope of '$\ltimes$'.

[15] This is also used in the theory of (linguistic) groups after Landman (1989).The function $\uparrow$ (written '{}') is also used by Oliver and Smiley (2016, §14.7). However, they assume it as a primitive and say "We do not need extensionality as an axiom, however, since it is implicit in the syntactic classification of {} as a function sign" (p. 268). On my end, starting from a relational symbol like '$\ltimes$' and justify the introduction of functional notation on the basis of its extensional behavior makes things more perspicuous and allows to avoid some of the oddities of their system (see §6.3.). Overall, unless one believes that all pluralities collapse into sets, like Florio and Linnebo (2021), choosing a relational collapse predicate over a functional one seems the best choice.

[16] While pluralities can be defined as any collection of the form above due to P-COMP, for obvious reasons sets cannot. Following textbooks presentations, the notation above generally defines a class, which is a set if it exists. See Kunen (2013).

[17] Some already in Burgess (2004) and Oliver and Smiley (2016).

is uniquely determined by the pluralities it is collapsed from. For instance, if the following '$\uparrow x$' exist, we have these equivalences:

$$\uparrow \|x : \phi(x)\| = \{x : \phi(x)\} \qquad \downarrow \{x : \phi(x)\} \approx \|x : \phi(x)\|$$
$$\uparrow\downarrow a = a \qquad \downarrow\uparrow aa \approx aa$$

Furthermore, 2.3(3) abuses notation from set-intersection to define the intersection between a set and a plurality. I take this to be uncontroversial as they are both collections subject to some kind of membership: since something can be both a member of a set and one among a certain plurality, there should be a collection gathering together what a set and a plural have in common. I take this to be a plurality because of the "*conceptual priority*" that characterizes pluralities with respect to sets in the Cantorian picture:[18] one could collapse a set-plural intersection to obtain the corresponding set: $\uparrow(a \cap aa) \coloneqq \{x : x \in a \wedge x \prec aa\}$. I also take intersections and unions between pluralities (for which I also abuse notation) to be uncontroversially defined just like their set-theoretic equivalents:

$$\bigcap xx \approx \downarrow \bigcap \uparrow xx \qquad \bigcap x = \uparrow \bigcap \downarrow x \qquad \bigcup xx \approx \downarrow \bigcup \uparrow xx \qquad \bigcup x = \uparrow \bigcup \downarrow x$$

Finally I also define the notation for the following special items:

Definition 2.4 (P-INDISC, EXT$_p$)**.**

(i) $ee :\approx \|x : x \neq x\|$ (i.e., the empty plurality)
(ii) $\uparrow ee = \emptyset$ and $\downarrow \emptyset \approx ee$
(iii) $\emptyset\emptyset :\approx \|x : x = \emptyset\|$ (i.e, the singleton plurality of the empty set)
(iv) $\uparrow \emptyset\emptyset = \{\emptyset\}$ and $\downarrow \{\emptyset\} \approx \emptyset\emptyset$

As mentioned above, the empty plurality has a somewhat controversial status (Oliver and Smiley, 2016). Here I take its existence to be uncontroversially granted simply by P-COMP. The reason is that our primary focus is the development of a conception of sets, so I am insensitive to whether or not the empty plurality has a correspondence in the natural language.[19] Similar considerations should dissolve also worries related to singleton pluralities. In any case, those worried by these pluralities may refer to the aforementioned Oliver and Smiley (2018) for a parallel project of a Cantorian "theory" of sets or may try to adapt this project of a "conception" of sets to Button (2021, Appendix A and B), who develops a level theory with urelements.

3. Iterations

Now that we have an idea of how pluralities and sets interact between each other and of how the first can, in principle, be collapsed into the second, we shall provide a structure to situate this intuition which is also our solution to the quest for consistent multiplicities. Then, we shall sketch a preliminary picture of how the process of plural-to-set abstraction can be placed within this iterative framework.

[18] Whether this priority can be turned into something like metaphysical dependence or some other hyperintensional notion is something I remain neutral on in the spirit of Incurvati's minimalism concerning the Iterative Conception (Incurvati, 2012, 2025).
[19] A more pressing issue concerning the empty plurality would concern its interaction with the "nothing over and above" conception advocated by Roberts (2022). I leave this discussion to further more philosophical work.

3.1. Iterating Sets

The structure is provided by the Iterative Conception of Set. Described in the most general way possible, this is a way of introducing the universe of sets in a more explicit and structural manner[20] and in more constructional terms,[21] that is, by directly regimenting a pre-theoretic story of how sets are generated through an iterated set-construction procedure:

> **Iterative Conception of Set**: sets are generated in layers. Each set lives at some layer. At every layer some set-generating operation outputs new sets by using as inputs sets found at previous layers. There are no further sets outside those generated in this way.

The operational talk in terms of generation follows an approach first presented by Forster (2008) and thoroughly developed by Button (2024) which places at the heart of the conception just the idea of *iterating some constructional procedure*. This contrasts the widespread impression that intrinsically ties the conception to the cumulative hierarchy and its set-generator: power-set. To use Forster words: "there is nothing in the idea of sets as conceived *iteratively* that says that there should be only one constructor [i.e., power-set]" (Forster, 2008, p. 99, emphasis in the text).[22]

The present project is not far from the original idea that the Iterative Conception is an explicit (i.e., non-recursive) way of describing and axiomatizing the cumulative hierarchy (see Montague, Scott, and Tarski, unpublished). The original shift is that it develops the conception as a way of reconciling (and making sense of) Cantor's original ideas concerning the process of set-formation with(in) the most popular development of the structure of the universe of set. In the end, the output resembles the hierarchy of sets under most respects, with the crucial twist that the set-generating operation is *plurality-to-set abstraction*, namely '$\ltimes$', following Forster, in focusing on a different constructional procedure.[23] This twofold nature is also justified by the fact that a Cantorian conception of sets as abstracted from pluralities is also (implicitly) present in authors explicitly engaged with the Iterative Conception. Here's how Dana Scott, introduces one the first axiomatizations of the conception: "But note that our original intuition of set is based on the idea of having *collections of already fixed objects*" (Scott, 1974, p. 207, my emphasis). Just as for Cantor's original remarks, it is most natural to interpret the "already fixed objects" as a plurality. The same idea is also present in George Boolos' famous paper which popularized the conception among philosophers:

> For when one is told that a set is a collection into a whole of definite elements of our thought, one thinks: Here are some things. Now we bind them up into a whole. *Now* we have a set. We don't suppose that what we come up with after combining

[20] The alternative approach, the one originally championed by Zermelo (1930), is the axiomatic one: start with some principles that describe the behavior of set and then simply generate the cumulative hierarchy (i.e., V) via a definition by transfinite recursion.

[21] Be careful not to confuse "constructional" with "constructive". The first term is borrowed from the literature on constructional ontologies and indicates a step-by-step construction of entities (sets in this case) through an iterative process. The second refers to Gödel's constructible universe, or L, which is a way of conceiving the Iterative Conception in predicative-friendly terms, but that is not the focus of this paper (see fn. 22).

[22] Furthermore, I remained silent on whether, at each layer, one finds all possible outputs the set-generator. This is in line with the most recent literature on the weak conception (Barton, 2024) that detaches the Iterative Conception from the cumulative hierarchy under another respect, namely that related to the *maximality* of the generative process. In other terms, besides there being more than just one constructor, there are also more ways to use each constructor, depending on whether one wishes to obtain all possible results of the application of that constructor at the very next layer. The most straightforward example is the one that compares V and L: the constructor is always power-set, but in the second case its strength is considerably weakened by yielding only definable subsets.

[23] I still move within a strong notion of the conception (see fn. 22) and interpret the process of set-formation as an abstraction from all possible pluralities found at some stage to all possible sets obtained from those pluralities.

> some elements into a whole could have been one of the very things we combined. (Boolos, 1971, p. 18, emphasis in the text)

In particular, the lasso metaphor that he attributes to Kripke in a footnote can be taken as a natural description of the process of plural-to-set abstraction: *forming a set is like throwing a lasso around some things, but you cannot throw a lasso if the things you are trying to group are not definite and fixed, on pain of paradox*. This is largely reminiscent of Cantor's talk of consistent and inconsistent multiplicities, which makes the decision of clarifying these notions within the conception rather natural.

3.2. Iterating Pluralities of Sets

A first suggestion on how the Iterative Conception can sharpen the concept of "consistent multiplicity" can already be found in one of its seminal formulations:

> This concept of set, however, according to which *a set is something obtainable from the integers (or some other well-defined objects) by iterated application of the operation "set of"*, not something obtained by dividing the totality of all existing things into two categories, has never led to any antinomy whatsoever; that is, the perfectly "naive" and uncritical working with this concept of set has so far proved completely self-consistent. (Gödel, 1947, pp. 518-519, my emphasis)

Here Gödel seems to be suggesting two things. The first is a rather straightforward and undisputed fact regarding the conception, namely its being a *combinatorial* rather than a *logical* view of the process of set-formation.[24] The second is that this combinatorial aspect can simply be instantiated by iterated application of a "set of" operation to some "well-defined objects".

Although the identification between this operation and my '$\ltimes$' may seem straightforward, we need a caveat first. As a matter of fact, Gödel (1951, fn. 5) specifies that by "set of" he means power-set, tying the Iterative Conception to this privileged constructor and its output (i.e., the universe V) as described by Forster in the previous passage. However, we should not feel discouraged by this. Our purpose is to give a precise definition of the notion of consistent multiplicity advanced by Cantor and to do so we are appealing to the tools offered by the Iterative Conception. It is then legitimate to adapt these tools to the present context, namely one where sets are generated by abstraction from pluralities. Moreover, pluralities perfectly capture the combinatorial aspect embodied by the Iterative Conception where sets are pinned down by their elements rather than an application condition.[25]

Therefore, as a first approximation, we can define as "consistent" all those multiplicities that are part of an iterative process of set-generation starting from a "safe" multiplicity, like the integers, or an empty multiplicity (i.e., ee). In what follows I shall show how this idea, in its latter instance,[26] can be sharpened further through a process that parallels the historical development of the Iterative Conception.

[24] See Incurvati (2020); Maddy (1983); Parsons (1974).

[25] See Florio and Linnebo (2021); Linnebo (2010).

[26] The former instance has been advocated by Oliver and Smiley (2016) in their attack on empty and singleton sets and defense of ur-elements in the context of the conception. Although I agree that the most faithful interpretation of Gödel's remarks would be taking the integers as an ur-element basis, it is also true that, as I noted above, Gödel also had in mind the iteration of a straightforward set-theoretic operation like power-set. For this reason, I think that if it is legitimate to distance myself from Gödel by iterating plural-to-set abstraction (as Oliver and Smiley did), then it is also legitimate to further distance myself by retrieving the pure universe of sets. Furthermore, it can also be argued that this is a fake distancing since Gödel showed no issues with the pure universe

3.3. The Plural Hierarchy

As a starting point, let's adapt Gödel's "set of" operation to the present setting where sets are generated from pluralities. We can move from the standard notion of power-set to a plural-based one, through the following operation, which pins down the sets generated from the subpluralities of a given plurality:

Definition 3.1 (P-IND). $x \trianglerighteq aa \leftrightarrow_{def} (\exists xx \preccurlyeq aa)xx \ltimes x$

This yields the "power-plurality" of a plurality 'xx' as the plurality of all the sets obtained by collapsing all the sub-pluralities of xx

Definition 3.2 (POWER PLURALITY). $\mathcal{S}(xx) :\approx \|x : x \trianglerighteq xx\|$

The equivalences deriving from Def. 2.3 yield a straightforward connection between this notion and the standard set-theoretic power-set:

Fact 3.3. For any x and xx: $\mathcal{S}(xx) \approx\!\downarrow \mathcal{P}(\uparrow xx)$ and $\mathcal{P}(x) =\!\uparrow \mathcal{S}(\downarrow x)$

For the sake of argument, let's put this definition at work in a context where ordinal indexing and transfinite recursion are primitively available, as in Pollard (1985), to obtain an approximate and preliminary sharpening of Gödel's remark. This serves the heuristic purpose of offering an intuitive grasp of the final picture, which will be perfected later, once we define our theory of levels.

Definition 3.4 (vv_α). $vv_0 \approx ee; \quad vv_{\alpha+1} \approx \mathcal{S}(vv_\alpha); \quad vv_\lambda \approx \bigcup_{\alpha<\lambda}(vv_\alpha),\ (\lambda \text{ limit})$

That is, iterating a plurally-interpreted "set-of" operation along the same process described by Gödel leads to a cumulative hierarchy of pluralities. This idea can be made more explicit if read in connection to the standard set-theoretic V_αs, again straightforward from Def. 2.3:

Fact 3.5. $vv_\alpha \approx\!\downarrow V_\alpha$ and $V_\alpha =\!\uparrow vv_\alpha$.

In other terms, we have defined a hierarchy of pluralities where each layer collapses into the corresponding one of the hierarchy of set, and vice-versa for the uncollapse. Imagine the two hierarchies sandwiched between one another, alternating a plural and a set-theoretic layer one at a time starting from the empty plurality.[27]

of sets, but instead was one of its most vocal advocates (see Kanamori, 2004). Therefore, I think that, rather than telling against my project, an appeal to Gödel tells against starting with ur-elements instead.

[27] This image can be made more perspicuous through a redefinition of both hierarchies in terms of the other, which can go in multiple ways:

$$
\begin{array}{lll}
vv_0 \approx\!\downarrow V_0; & vv_{\alpha+1} \approx \mathcal{S}(\downarrow V_\alpha); & vv_\lambda \approx \bigcup\limits_{\alpha<\lambda} (\downarrow V_\alpha),\ (\lambda \text{ limit}) \\
V_0 =\!\uparrow vv_0; & V_{\alpha+1} = \mathcal{P}(\uparrow vv_\alpha); & V_\lambda = \bigcup\limits_{\alpha<\lambda} (\uparrow vv_\alpha),\ (\lambda \text{ limit}) \\
vv_0 \approx\!\downarrow V_0; & vv_{\alpha+1} \approx\!\downarrow \mathcal{P}(V_\alpha); & vv_\lambda \approx\!\downarrow \bigcup\limits_{\alpha<\lambda} (V_\alpha),\ (\lambda \text{ limit}) \\
V_0 =\!\uparrow vv_0; & V_{\alpha+1} =\!\uparrow \mathcal{S}(vv_\alpha); & V_\lambda =\!\uparrow \bigcup\limits_{\alpha<\lambda} (vv_\alpha),\ (\lambda \text{ limit}) \\
vv_0 \approx ee; & vv_{\alpha+1} \approx\!\downarrow \mathcal{P}(\uparrow vv_\alpha); & vv_\lambda \approx\!\downarrow \bigcup\limits_{\alpha<\lambda} (\uparrow vv_\alpha),\ (\lambda \text{ limit}) \\
V_0 = \emptyset; & V_{\alpha+1} =\!\uparrow \mathcal{S}(\downarrow V_\alpha); & V_\lambda =\!\uparrow \bigcup\limits_{\alpha<\lambda} (\downarrow V_\alpha),\ (\lambda \text{ limit})
\end{array}
$$

Alternating between up and down arrows should give an idea of how one can move between plural and set layers.

This way of conceiving the process of set construction basically condensates what we could label the "pre-history" of the Iterative Conception, namely the thirty years that went from the discovery of the paradoxes to Zermelo's formulation of the Cumulative Hierarchy. If we indulge for a moment in the practice of counterfactual history, it makes sense to imagine that the early set theorists would have ended up with the vv_α had they given more space to Cantor's intuition that the process of set formation is a process of abstraction from pluralities. Just like the set-theoretic hierarchy grounds a paradox free (i.e., consistent) process of set formation, we can take the above hierarchy of pluralities to ground a notion of "consistent multiplicity" in the same way. A *consistent multiplicity is any plurality that occurs at some level of the cumulative hierarchy of plurals*, that is:

Definition 3.6 (CONSISTENT MULTIPLICITY)**.** $\mathfrak{C}(xx) \leftrightarrow_{def} \exists\alpha(xx \preccurlyeq vv_\alpha)$

The Cantorian nature of this definition rests also on the fact that there is still room for inconsistent multiplicities. That is, we are not endorsing something like the following principle: $\forall xx \exists\alpha(xx \preccurlyeq vv_\alpha)$. This would be straight away inconsistent with P-COMP since pluralities like the pluralities of all sets do not appear at some level of the plural hierarchy. This principle can be accepted if we restrict comprehension. In fact, Florio and Linnebo (2021) commit to an equivalent axiom formulated in terms of an induction scheme when they prepare the setting for their Critical Plural Logic to derive ZF. I, on the other hand, am committed to a weaker principle, which is sufficient for the purposes of the present theory: $\forall xx(\mathfrak{C}(xx) \rightarrow \exists\alpha(xx \preccurlyeq vv_\alpha))$, i.e., *every consistent multiplicity appears at some level of the plural hierarchy*. This is just a consequence of Def. 3.6 and flags the fact that *the plural hierarchy pins down exactly the consistent multiplicities*, which is a way of sharpening Cantor's claim via Gödel's observation. An analogous formulation, which is a crucial theorem of the theory presented in next section, is that the plural hierarchy is well-founded.

That said, however, the cumulative hierarchy is not all there is to the Iterative Conception. On the contrary, the conception as we know it resulted from the hierarchy once a "notable inversion" has been enacted: "In a *notable inversion*, what has come to be regarded as the underlying iterative conception became a heuristic for motivating the axioms of set theory generally" (Kanamori, 2004, p. 521, my emphasis). That is, once defined via the means of axiomatic set theory, the hierarchy motivates a notion of layer that can be explicitly and non-recursively defined, quantified over and axiomatized. This is, in a nutshell, the inversion brought to the table by axiomatic approaches to the conception. Therefore, in a parallel move, I now proceed to offer an analogous regimentation of what the Iterative Conception looks like if axiomatized against the background of the vv_α. Once again, we can leverage on the work already done in the set-theoretic case in the past 90 years, which leads to the formulation of a theory of plural levels.

4. Plural Level Theory

Plural Level Theory (PLT), as an axiomatization of the Iterative Conception, is a plural reworking of the Level Theory (LT) advanced by Button (2021).

4.1. Intuitive Formulation

Let's start with a pre-theoretic story that captures its general idea:

The Plural Story: sets are generated in layers. *Some things are a layer* just in case they are *all possible sets* of sets found at previous layers. Each set lives at some layer. There are no further sets outside those generated in this way.

PLT takes the plural talk of the story at face value and makes it explicit through an axiomatization. That is, it interprets "all possible sets of sets" as a plurality of sets and regiments this notion of plural layer through an explicit definition.[28]

Before proceeding with the formalization, it is preferable to sketch an informal characterization of how I want the hierarchy of plural levels to behave. In this way we can recognize a pattern that should then lead to a suitable regimentation. The strategy is to parallel the Button-Potter approach of alternating between *histories* (accumulations of levels) and *levels*. Starting from the empty set,[29] we generate the first level, accumulate it in the next history, generate the second level, accumulate all the previous levels in another history, generate a level and so on. As stated in Button and Walsh (2018, 8c), we can generally portray levels as the V_αs and histories as sets of the following form: $\{V_\gamma : \alpha < \gamma\}$.[30]

On our end, representing plural levels as the unbracketing of the set-theoretic levels, that is, our vv_αs is rather straightforward. However, two strategies are now available for what concerns the notion of a plural history, depending on whether we lean more towards being faithful to our set-theoretic aims or towards being "plural-purists" so to speak. Here I focus on the first case, and I take both levels and histories to be the "unbracketing" of Button's set-theoretic notions. Therefore, combining the two heuristics above, plural histories will simply be the following pluralities: $\| \uparrow vv_\gamma : \gamma < \alpha \|$. Here plural histories gather the $\uparrow vv_\alpha$s (i.e., the V_αs), into pluralities. In the second case, plural levels are accumulated into histories while remaining pluralities without being collapsed. To do so it seems that we are forced to appeal to super-pluralities because, being an entity of one order higher than pluralities, they can keep track of the difference between, for instance, vv_0, vv_1 and vv_2.[31] However, since this a rather controversial topic that deserves its own discussion,[32] I leave the formulation of a Super-plural Level Theory to a separate work.

[28] Those familiar with the axiomatizations of the conception may have noted that I intentionally skip one passage from Button's blueprint, namely the formulation of a theory of (plural) *stages* before that of (plural) *levels*. After Boolos (1971, 1989); Kreisel (1965), these are *sui generis* entities primitively quantified over, which seems the most natural first move when axiomatizing the pre-theoretical story. Although parting with the level-theoretic tradition (Montague, Scott, and Tarski, unpublished; Potter, 1990, 2004), Button still has to face the approach popularized by Boolos. He first develops a stage theory and then shows that it makes the same set-theoretic claims of LT, arguing for the latter through an economy-of-primitives argument. Here I can avoid the confrontation with a hypothetic plural-stage-theoretic approach simply because there is none. Even if there was one, this would be an over-complication with no significant advantages or justifications since bare plural quantification because already is a natural way to formalize the plural story. I also subscribe to the reasons advanced by Button in moving from stages to levels: just like he aims for a purely set-theoretical foundation, here I aim for a "purely plural" development of the conception. Moreover, given the tight connection between PLT and LT (see below), I take my case to be analogous in the sense that, if developed, a plural stage theory could easily be made set-theoretically equivalent to PLT.

[29] Which is trivially a history, see Potter (2004).

[30] Note that this equivalence is not just a useful heurstic to explain what levels and histories are in terms of something familiar, but an actual result proved by Button and Walsh. That is, one can prove that the levels are exactly the V_αs and vice-versa.

[31] Take the accumulation of vv_0, vv_1 and vv_2 in the plural hierarchy outlined in the previous section. If we take this to be just a plurality we would have the following result, which loses track of two of the three levels: $\text{HIST}_3 \approx vv_0 + vv_1 + vv_2 \approx ee + \emptyset\emptyset + \emptyset, \{\emptyset\} \approx \emptyset, \{\emptyset\}$. In other terms, since plurals are nothing over and above their members (Roberts, 2022), accumulating those three pluralities together would amount to have a plurality where the empty plurality disappears since it does not contain anything, while the empty set is counted twice and, just as in the case of sets, we don't count the members of a plurality more than once. As a result, in this case, we lose track of vv_0 and vv_1 in the history, hence the appeal to super-pluralities.

[32] See Florio and Linnebo (2021, ch. 9), Linnebo and Nicolas (2008); Nicolas and Payton (2025); Payton (2025).

4.2. Formal Regimentation

Now that we have an intuitive representation of the picture advanced by PLT, we can move to the definition of the relevant concepts. Let's start from the definition of an operation that Button (2021), following Montague, Scott, and Tarski (unpublished); Scott (1974), labels *potentiation*:[33]

Definition 4.1 (PLURAL POTENTIATION)**.** For any aa, let aa's plural potentiation be $\P\P(aa) :\approx \|x : x \blacktriangleright aa\|$.[34]

In other terms, a plural potentiation of a plurality aa is a plurality consisting of all the subsets of members of aa. This mirrors Button's potentiation as the set of all the subsets of the elements of the given set, that is, a super-transitive closure (closure under subset, i.e., $\P(a) := \{x : x \rhd a\}$). So the plural definition coincides with that of a super-transitive plurality. Moreover, it also preserves the "conceptual connection" with power-set observed by Button (2021, p. 439): if the singular potentiation of a singleton set is equivalent to the *power-set* of its unique member ($\P(\{a\}) = \mathcal{P}(a)$), the plural potentiation of a singleton plurality corresponds to the *power-plurality* of its sole member, i.e., the plurality that then collapses in the power-set of its member.

Fact 4.2. If aa is the singleton plurality of the set a, then, if it exists, $\uparrow \P\P(aa) = \mathcal{P}(a)$.

Moreover, since we defined the a new operation, this correspondence can be also tracked town in terms of a power plurality:

Fact 4.3. If aa is the singleton plurality of the set a, then $\P\P(aa) = \mathcal{S}(\downarrow a)$.

Next, we define a *plural history*:

Definition 4.4 (PLURAL HISTORY)**.** $\textsc{Hist}(uu) \leftrightarrow_{def} \forall x \prec uu(x =\uparrow \P\P(x \cap uu))$

Here the definition twists Button's, whose histories are sets whose members are the potentiation of the intersection between the set and their members. The reason is the plural setting: since our aim is to build a plural hierarchy of sets, we need a place to introduce new sets, that is, to tap the "wand" '$\uparrow$' (or '$\ltimes$') to use Forster's and Button's terminology, otherwise we would move in a tiny circle and the iterative process would never kick in. For this reason, a plural history is a plurality whose members are sets collapsed from the plural potentiation of the plural-set intersection between the history itself and its members.

Finally, we define the core concept of a *plural level*:

Definition 4.5 (PLURAL LEVEL)**.** $\textsc{Lev}(ss) \leftrightarrow_{def} \exists uu(\textsc{Hist}(uu) \wedge ss \approx \P\P(uu))$

In other terms, just as in LT, levels are (plural) potentiations of histories, which are initial sequences of levels.

Lastly, we need an additional notion with respect to Button's LT:

Definition 4.6 (BOUNDED LEVEL)**.** $\textsc{Lev}_{\beta}(ss) \leftrightarrow_{def} \textsc{Lev}(ss) \wedge (\exists tt : \textsc{Lev}) ss \nprec tt$

The reason for this is that, as we shall see in §5.2., the theory sees the first-order domain as a plural level and thus P-COLLAPSE below stated for levels in general would force us to collapse not only the first domain, but also all of its sub-pluralities (i.e., subsets). Since the first-order domain does not contain any of these sets, this means that the axiom thus stated turns out false.

[33] The symbol '¶' was first introduced in Montague, Scott, and Tarski (unpublished) and comes back in Scott (1974).

[34] Contrary to Button's LT we don't add the qualification "if it exists" because P-COMP is enough to grant the existence of this (impredicative) plurality.

This is a basic consequence of the fact that the theory we are about to state actually is a theory of levels with one level of classes (see §5.5.), as anticipated at the end of Montague, Scott, and Tarski (unpublished, §22). That is, all the levels are bounded except the last one, which is the single unbounded level that covers the whole underlying plural hierarchy.

We are now ready to state Plural Level Theory.

Definition 4.7. PLT is the theory formulated in $\mathcal{L}_{\prec,\ltimes}$ with the following non-logical axioms: EXT_p +

$$\forall xx(\exists ss(\text{LEV}_\beta(ss) \wedge xx \preccurlyeq ss) \rightarrow \exists a(xx \ltimes a)) \qquad \text{(P-COLLAPSE)}$$

$$\forall a \exists ss(\text{LEV}(ss) \wedge a \prec ss) \qquad \text{(P-STRAT)}$$

P-STRAT is a plural analogue of the stratification principle that characterizes all axiomatizations of the conception: *every set lives at some plural level*, or *the plural level-hierarchy covers the whole universe of sets*. P-COLLAPSE is our "main engine of set production" (Yablo, 2006) and provides the final sharpening of Cantor's consistent multiplicities. The following definition parallels Def. 3.6 on the other side of Kanamori's notable inversion, offering an explicit characterization that does not hinge on ordinal primitives:

Definition 4.8 (CONSISTENT MULTIPLICITIES$^{\text{L}}$)**.** $\mathfrak{C}^{\text{L}}(xx) \leftrightarrow_{def} (\exists ss : \text{LEV}_\beta)xx \preccurlyeq ss$

That is, *consistent multiplicities* just are *pluralities bounded by bounded-levels*. Here we can interpret the two main axioms of PLT as providing the "*if*" (P-COLLAPSE) and the "*only if*" (P-STRAT) part of the Cantorian claim that a multiplicity forms a set if and only if it is consistent. Moreover, the left-to-right direction provides a parallel improvement of the principle mentioned at the end of §3.3.: $\forall xx(\mathfrak{C}^{\text{L}}(xx) \rightarrow (\exists ss : \text{LEV}_\beta)xx \preccurlyeq ss)$, i.e., *every consistent multiplicity appears at some (bounded) plural level*, or *the hierarchy of (bounded) plural levels pins down exactly the consistent multiplicities*. In fact, PLT can do more and prove that (bounded) levels are well ordered by plural inclusion '$\preccurlyeq$'. This means that the (bounded) levels really pin down all consistent multiplicities as generating the well-founded hierarchy of sets.[35]

5. Some results in PLT

Now that we have stated the theory, let's survey some of the most interesting results to be obtained in it.

5.1. The Fundamental Theorem of PLT

Let's start from what Button labels the *fundamental theorem*, namely the result I just mentioned about the well-ordering of the levels:[36]

Theorem 5.1. *Plural levels are well ordered by* $\preccurlyeq$.

[35] To avoid specifying it every time, from here on I will talk of levels in general and take it to be clear from context, if not specified otherwise, when I mean "bounded level".

[36] I am especially grateful to Tim Button for his helpful insights on the general strategy for the proof and, in particular, for a Lemma that got me stuck for months (see fn. 66).

The reason for the label is that Th. 5.1 yields both Foundation and $\in$-induction.[37] As remarked by Wang: "The axiom of foundation sharpens the concept of iteration" (Wang, 1974, p. 216). This becomes evident especially after Scott's first proof of this theorem, a result later celebrated by both Boolos (1984, 1989) and Button. The derivation of Foundation is the real hallmark of the Iterative Conception, which provides a uniform structure within which sets are arranged. The remarkable fact about the axiomatizations of the conception is that they induce this fundamental structure by a simple and explicit description of the layers of the hierarchy. That is, instead of assuming the principle at the outset, theories of levels incorporate it in the core definition of a level. Moreover, going back to Forster's description of the conception, this result shows that the final set-up really is uniform, in that the hierarchy turns out to be well-founded even though the set-generator is changed. To use Scott's own words: "This at first surprising result shows how little choice there is in setting up the type hierarchy" (Scott, 1974, p. 210).

The details of the proof of Theorem 5.1 are a bit tedious, although they differ a bit from Button's proof for reasons of bookkeeping. I therefore defer them to Appendix 1.1.. Still, there are two interesting details about the proof. First, it can go through in a theory weaker than PLT, where P-COMP, is restricted:

Definition 5.2. PLT$^-$ is PLT with P-COMP substituted by

$$\forall xx(\exists yy\forall u(u \prec yy \leftrightarrow (u \prec xx \wedge \phi(u)))) \qquad \text{(P-SEP)}$$

This is relevant because it makes the fundamental theorem available also to those scholars, like Florio and Linnebo (2021), that propose a weaker logic for plurals. Second, it is the collapse predicate '$\ltimes$' that carries over the heavy work of providing a satisfactory notion of minimality (see fn. 37), which offers an interesting insight in the kind of cross-type phenomena at the core of this approach.

5.2. Models of PLT

Let's consider the models of PLT. After Boolos (1985), these are interpreted as the usual models for full second order logic, where the power-set of the first-order domain acts as the domain for the second-order (i.e., plural) variables.[38]

Fact 5.3. $\langle V_\alpha, \mathcal{P}(V_\alpha), \ltimes\rangle \models PLT$ for all $\alpha > 0$

This means that PLT is maximally neutral with respect to the pluralities, and hence the sets, it generates. Although this may seem as a shortcoming of the theory, it is not. The reason is that this maximal level of neutrality, or, better, generality, coheres with the original spirit with which these theories of levels where originally formulated:[39]

> We could say that in the discussion above we were interested in the set-theoretical sentences true in relational systems of the form $\langle R(\alpha), \in_{R(\alpha)}\rangle$, for various ordinals α. [...] The problem we now wish to discuss is *which sentences are true in all*

[37] To be precise the result that sanctions both are two similar statements involving '$\ltimes$' and needed to derive the theorem. However, since the collapse predicate would act only as a "trans-type" kind of well-ordering, I focus on the more familiar "trans-type" notion. See Appendix 1.1..

[38] As noted by Burgess (2004, §9), there is little point in complaining against these "official" models in favor of "more natural" purely plural models, since Boolos showed that the two ways of conceiving models for plural logic are equivalent given P-COMP and the Axiom of Separation for sets. Therefore, despite the naturalness of plural models, for reasons of readability and ease of exposition, I use the official set-theoretic models with no loss for the Cantorian nature of the project.

[39] The same remarks are found in published form in Montague (1965).

> *systems $R(\alpha)$, where α is a non-zero ordinal*. The usual situation involving Gödel's Incompleteness Theorem arises here, and it can be shown that the set of sentences common to all systems $R(\alpha)$ cannot be axiomatized by a recursive set of axioms. Even so, we can present *a short and simple set of axioms adequate for the main properties of these systems*. (Montague, Scott, and Tarski, unpublished, p. 160, my emphasis)

That is, the mathematical purpose of a theory of levels is to provide a maximally general framework with respect to the set-theoretic landscape it describes. We can thus interpret PLT as a plural realization of the original project also embodied by Button's LT.

Considering models for PLT also allows us to justify the restriction on P-COLLAPSE. Consider the model of PLT for $\alpha = \omega$, which can be represented as follows:

$$V_\omega = \{\underbrace{\ldots\ldots\ldots}\}$$

$$\mathcal{P}(V_\omega) = \{\overbrace{\ldots\ldots\ldots}^{\asymp} \mid \ldots\ldots\ldots \mid V_\omega\}$$

The brackets and the pipes serve to visualize and separate the "collapsable" part of the second-order domain from its "uncollapsable" part and also from its maximal element, namely the whole first-order domain. This is a nice example to illustrate the restriction on P-COLLAPSE: if we had that any level-bounded plurality whatsoever collapses into a set, then the whole uncollapsable part of the second-order domain (and the first-order domain itself!) would collapse into sets of the first order domain, which clearly cannot be the case. Moreover, it also that there always is a last plural level (not necessarily a last bounded), namely the first-order domain itself, which PLT sees as a level or, more precisely, as its sole unbounded level.[40]

5.3. PLT and Set Theory

It is now time to check how PLT interacts with a theory of sets like Zermelo-Fraenkel set theory (ZF). This is the main philosophical reason behind the axiomatizations of the conception: providing a more natural and intuitive justification of the axioms of set theory. The same should be said for our Cantorian theory of levels, which can be reconnected to the developments of axiomatic set theory that followed Cantor. To justify ZF, however, we are forced to augment PLT with suitable bolt-ons to obtain those principles that the theory alone cannot sanction. For the moment, this is all the theory can do:

Proposition 5.4. $PLT \vdash$ SEPARATION, UNION, FOUNDATION

The first two additional principles are a straightforward adaptation from Button (2021), where all plural variables stand for bounded plural levels:

$$\forall ss \exists tt (ss \preccurlyeq tt) \qquad \text{(P-END)}$$

$$\exists ss [(\exists qq \preccurlyeq ss)(\forall qq \preccurlyeq ss)(\exists rr (qq \preccurlyeq rr \preccurlyeq ss))] \qquad \text{(P-INFINITY)}$$

Namely, there is no last level and there is an infinite level, which yields the following result as for LT:

[40] In algebraic terms this is a simple consequence of the fact that, since the standard second-order domain is always the power-set of the first-order domain it always induces an algebra, namely the power-set algebra, whose top element is the first-order domain itself.

Proposition 5.5.

(i) $PLT +$ P-END $\vdash$ PAIRING, POWER-SET
(ii) $PLT +$ P-END $+$ P-INFINITY $\vdash$ INFINITY$_Z$[41]
(iii) $PLT +$ P-END $+ \neg$P-INFINITY $\vdash ZF_{\text{FIN}}$[42]

The second principle is a bit more complex as it derives Replacement, a notoriously troublesome principle in the context of the Iterative Conception.[43] In LT this is captured through full second-order means. That is, for any functional F:

$$\forall F \forall a (\exists s : \text{LEV})(\forall x \in a) Fx \subseteq s \qquad \text{(UNBOUNDED)}$$

Its intended meaning is that the the hierarchy of levels is so tall that no set can be mapped unboundedly into it. Despite the fact that we have full P-COMP and so the plural reading of full second-order logic from Boolos (1984, 1985), we cannot go for a straightforward translation of Button's principle. Simply translating 'Fx' as '$x \prec xx$' would not be enough because we would still be lacking a mapping from a to xx. That is, it is true that P-COMP allows us to define arbitrarily big pluralities that may even exceed the hierarchy of bounded plural levels. However, '$x \prec xx$' does not capture the same meaning of 'Fx' even though they are both generated by an unrestricted comprehension principle. The reason is that, being functional, that the latter already embodies the concept of a mapping.

A first reply to this issue would be bite the bullet and go schematic and re-state Button's principle for each functional formula 'ϕ'. However, I do not think that this is satisfactory, precisely because of Boolos' reading of plural-logic, which places it on a par with second-order quantification while avoiding its shortcomings. In other words, since the logic of plurals does not lack expressive resources, there must be a way to make sense of quantification over functions in this context. This solution, it turns out, is simply to properly exploit the power of P-COMP to single out the notions of a functional plurality 'ff^a' over a set 'a' and its plural image '$ff^{[a]}$'. This is allowed simply by the above result concerning the derivation of the fragment of ZF and by the comprehension principle. However, since the details are more a matter of bookkeeping rather than being really interesting, I leave them to Appendix 1.2.. In the end, the principle looks like this:

$$\forall a (\forall xx \approx ff^a)(\exists ss : \text{LEV}_\beta) ff^{[a]} \preccurlyeq ss \qquad \text{(P-UNBOUNDED)}$$

In other words, the plural image of any functional plurality over a set is consistent and can thus form a set.[44]

The final result is again straightforward after LT's analogous theorem:

Theorem 5.6. *Let* PLT^+ *be* $PLT +$ P-INFINITY $+$ P-UNBOUNDED. *Then* $PLT^+ \vdash ZF$

[41] There is a set containing the empty set and closed under successor.
[42] See Button (2021, fn. 27).
[43] See Boolos (1971, 1989); Potter (1990, 2004).
[44] Moreover, note that an alternative route to obtain both Infinity and Replacement is the one through Reflection Principles first pursued by Scott (1974) in the context of the Iterative Conception. However, since this approach has already been pursued by Burgess (2004), who interprets it as way to make sense of a limitation of size view, I defer the discussion of this principles to the next section.

5.4. PLT vs. LT

The reason why the above results are straightforward is due to the relation between PLT and its set-theoretic second-order cousin LT:[45] the two theories are synonymous in the sense of being definitionally equivalent. Once again thanks to Boolos (1984, 1985), this result is straightforward: one just interprets the second-order variables of LT along the plural interpretation (see next section for the primitive predicate '$\ltimes$').[46] Although this equivalence may be interpreted as a trivialization of PLT as a mere plural copy of LT, I think this thought is misguided.

For one thing, PLT originates from a completely different standpoint which makes my project and Button's quite different. On the one hand, I want to provide a theory of Cantorian "consistency" (or co-existence, see §6.2.) for pluralities that sharpens Cantor's idea that sets are obtained by collapsing pluralities. To do so, I employ the tools provided by the Iterative Conception since this route to consistent multiplicities had not been explored yet. On the other hand, Button moves already within the framework of the conception and his appeal to full second-order logic is not substantial, but only instrumental to obtain the quasi-categoricity theorem. That is, LT does not put the same weight as PLT on higher-order variables simply because it already is a theory of sets that could work also if formulated in first-order terms. On the contrary, PLT is grounded on a substantial appeal to pluralities to generate sets along an iterative process which is instrumental to grant the consistency of the plural-to-set abstraction.[47] On top of this, I think that this is a rather comforting result. The reason is that, no matter how one finds the plural-talk or the collapse idea extravagant and exotic, pushing back against PLT becomes quite hard if one is also provided with a "safe" and "familiar" retreat into a standard second-order theory of sets like LT.

That said, since we mentioned that Button's second-order formulation of LT is instrumental for obtaining quasi-categoricity, it should not be surprising that the same result carries over also to PLT. This is another crucial result in these theories, along with the well-ordering of the levels, first proven by Montague, Scott, and Tarski (unpublished) and later perfected in Button and Walsh (2018), who also add the qualification of internal categoricity.[48]

5.5. Classes in PLT

One of the chief advantages of retrieving Cantor's picture through PLT is that, together with a conception of sets, it also legitimizes the approach to proper classes presented in the seminal Uzquiano (2003). This has two positive consequences. First, Uzquiano's main point is that plural logic make sense of class talk, especially in its impredicative form as axiomatized in Morse-Kelley class theory (MK).[49] This is not only useful, but sometimes indispensable to formulate

[45] LT is natively a second order theory so we do not need to flag it. It's first order counterpart is tagged as LT_1 instead.

[46] Alternatively, we could plug in a collapsing predicate for LT's second order variables, modify it along the same lines of PLT and the result would be even more perspicuous.

[47] This is also why the suggestion from fn. 46 of equipping LT with a collapsing predicate to make the equivalence with PLT more perspicuous is not so trivial. It would rather carry over a substantial commitment to a certain view regarding the process of set formation.

[48] Internal categoricity being the object-language claim, derivable from second-order logic, that there is a relation that can internally code (or mimic) the behavior of an isomorphism between two models. See Button (2021, §6) who also strengthens to full internal categoricity. I redirect the reader to the aforementioned texts for the proofs since here there are no relevant differences between the two cases.

[49] Here we shall be careful to interpret MK without a principle of limitation of size *à la* von Neumann (1925, 1928), i.e., *all proper classes have the same size as the universe*. The reason is that PLT is not equipped with a choice-like axiom in a plural guise after Burgess (2004). Here I agree with Boolos' on AC not being obviously entailed by the Iterative Conception, while disagreeing on the same view concerning Replacement and being closer to Shoenfield (1967, 1977). However, caution should be used even if we accept Burgess' Plural Choice since this must be equivalent to Global Choice for the theory to be equivalent to MK plus limitation of size (see Fraenkel, Bar-Hillel, and Lévy, 1973). This is not immediately obvious since Burgess claims that his axiom is equivalent to

certain set-theoretic statements like large cardinal hypotheses.[50] Second, this completes the interpretation of Cantor's passages, where both consistent and inconsistent multiplicities are part of the picture. In PLT level-bound pluralities collapse into sets and unbounded pluralities do not. In class terms these correspond, respectively, to sets (*level-bound pluralities*) and proper classes (*level-unbounded pluralities*), which naturally matches Cantor's distinction between consistent and inconsistent multiplicities (see §6.1.).

Furthermore, PLT provides a natural interpretation of Uzquiano's notion of "correspondence" between a class and a set (p. 74), namely our primitive predicate '$\ltimes$'. More specifically, had we taken the usual set-membership predicate '$\in$' as primitive rather than '$\ltimes$', we could have defined the latter in the same way Uzquiano defines the correspondence between a set and a class: $xx \ltimes x \leftrightarrow_{def} \forall y(y \in x \leftrightarrow y \prec xx)$.[51] In particular, the current approach is enough to provide the plural reading of two-sorted MK mentioned by Uzquiano.[52] This further yields a perspicuous reading of Separation and Replacement as full axioms rather than schemas and further clarifies the seemingly weird notion of "functional plurality" invoked above simply as a functional class.

Moreover, this has the remarkable consequence of establishing PLT as a realization of the rank-free theory with classes originally proposed by Montague, Scott, and Tarski (unpublished), which is the first complete instance of a theory of levels. That is, after proposing their theory for sets (of which Button's LT is the latest and most developed descendent), Montague, Scott and Tarski speculate that one could provide a similar theory of levels with n levels of classes, where $n = 1$ would be equivalent to MK.[53] However, as for sets, their realization is rather convoluted and not so easy to follow. Therefore, in light of the above remarks concerning the reading of plurals as classes and PLT's unique unbounded level, i.e., class-level, we can say that our theory realizes the aforementioned project in a more perspicuous way, just as Button's LT is a more straightforward realization of their original theory of levels.

6. Other Approaches

PLT is the alone in trying to make sense of Cantor's remarks concerning pluralities and sets. In this section we examine rival approaches and argue for PLT as making better sense of set formation as a process of plural-to-set abstraction.

6.1. Limitation of Size

As mentioned in the opening, the project of making sense of Cantor's remarks through to the Iterative Conception of Set is compatible with Cantor's original ideas although it is probably

standard set-theoretic AC. However, since the proof happens against the background of a theory which includes a plural (read "second-order") reflection principles, there is still room to believe the axiom to be a version of global choice. This is enforced by the plural reading of classes and the fact that Burgess' axiom states the existence of a choice function over pluralities *simpliciter*, where these are generated by P-COMP. This makes the case for plural choice as a logical principle a bit troublesome, unless one subscribes to a limitation of size view of the process of set formation as Burgess does (see §6.1.).

[50] Although other strategies may be available, such as going property-theoretic or taking class-talk at face value, the plural strategy seems the most promising and less problematic, especially after Boolos' work. Thanks to an anonymous referee for pointing out these alternatives to me.

[51] This is also the way in which LT would interpret PLT.

[52] See Uzquiano (2003, p. 78). While the plural interpretation of two-sorted MK is immediate (set variables are mapped to sets and class variables are mapped to pluralities), one-sorted MK has to be translated to the former for the result to go through. While the transfer is almost trivial, as noted by Uzquiano himself, given the categoricity of PLT, the correspondence with one-sorted MK may not be so trivial. I leave this to further work.

[53] This would seem to go against the usual view, also endorsed by Uzquiano (2003), that classes do not iterate. However, as the authors remark: "Further, the theories obtained by setting n [i..e, the number of iteration of classes] equal to 2 or 3, though unknown in the literature, can be quite useful for the development of certain branches of mathematics, for instance, parts of algebra and model theory". (Montague, Scott, and Tarski, unpublished, p. 178). It is an interesting matter for future work to understand what commitments (ideological or ontological) this further levels of classes force us into, especially in light of the plural reading of classes and of the recent debate on super-pluralities and its possible elimination through plural cover predicates (see fn. 32).

not faithful to its original intents (Ferreirós, 2007). As thoroughly documented by Hallett (1984) it makes more sense to attribute him a *Limitation of Size* Conception, traditionally opposed to the Iterative Conception.[54] A nice way of characterizing the two approaches comes from Uzquiano (2003) by adapting his talk of "from above" and "from below" principles (pp. 70-71). Starting from the Iterative Conception, the whole idea is to axiomatize a constructional procedure that builds sets *from below* by iterating the application of some operation. Then, one may ask if there are collections so big that they escape this process of stratification. In a sense, these would be "too big from below" and PLT characterizes them as the opposite of "consistent multiplicity":

Definition 6.1 (INCONSISTENT MULT.)**.** $\neg\mathfrak{C}^{\mathrm{L}}(xx) \leftrightarrow \neg(\exists ss : \mathrm{LEV}_{\beta})xx \preccurlyeq ss$

That is, PLT describes inconsistent multiplicities as those pluralities that are level *unbounded*, i.e., that cannot be built "from below" by iteration or that are "un-stratifiable" so to speak: *if, running through the levels (from below), none of them binds* xx*, then* xx *is inconsistent*, i.e., it is a proper class.

On the other hand, Limitation of Size determines *from above* what is a set or not, depending on whether a collection can or cannot be put into one-one correspondence with collections, like the universe of sets (von Neumann, see fn. 49) or the ordinals (Cantor, see Hallett, 1984). Here the idea is the opposite, namely "too big from above". These collections are not too big because they escape a process of stratification that starts from below, but are rather given in advance, i.e., from above, as too big and then are used to characterize sets as small collections (so perhaps here the correct label is "small from above"). In a sense, the two conceptions reverse each other's order of explanation, focusing either on sets or classes first: the Iterative Conception builds sets from below and then says that classes escape the stratified structure of the set-universe; Limitation of Size first assumes classes as big collections and then uses them to pin down sets as small collections. In other words, the contrast is a matter of *conceptual priority between sets and classes*.

According to Burgess (2004), the idea of limitation of size is also captured by reflection principles, originally described as from above limitations by Uzquiano: properties of the entire universe are reflected down to initial segments of the hierarchy of sets. It is on these principles that Burgess builds his theory BB (Boolos-Bernays) as a plural theory of sets that should capture the Limitation of Size Conception.[55] In particular, the plural reading of classes is essential for Burgess as it enables a principle of class-reflection *à la* Bernays (hence the name) suitable to derive not only the axioms of Infinity and Replacement, but even large cardinal principles up to Mahlo. The principle so interpreted also makes sense of the Cantorian distinction by defining *inconsistent multiplicities* as those pluralities such that "... any statement Φ that holds of them continues to hold if reinterpreted to be not about all of them but just about some of them, few enough to form a set" (Burgess, 2004, p. 205).

So interpreted, I do not think that Burgess' characterization is a satisfactory account of the Cantorian view. The reason is that, it already presupposes the notion of set and so it does not go much further than Cantor in simply stating that consistent multiplicities are those pluralities that do (and inconsistent those that do not) form sets. On the contrary, PLT explains the distinction on the basis of a proper description of the process of set formation as arising from plural abstraction. That is, consistent multiplicities are not just pluralities that form sets, but rather pluralities that are bounded by the (explicitly defined) levels of the cumulative

[54] see Wang (1974, ch. 6) and Hallett's critique of the Iterative Conception in his §6.1.
[55] See also Pollard (1996) on this same link between pluralities and limitation of size.

hierarchy, while inconsistent multiplicities are those that escape these boundaries. To put it in more perspicuous terms: *PLT first explains the notion of (in)consistency as level (un)boundedness and then says that consistent multiplicities are those that form set, it does not just say that some multiplicities are consistent because they are those that form sets*. While I do not take this to be a definitive argument in favor of the Iterative Conception over Limitation of Size, I nonetheless interpret it as explicitly favoring the former to make sense of Cantor's ideas concerning the formation of sets and the contrast with classes.

On top of this, there are two further considerations. First, despite clearly being "from above", reflection principles are not an exclusive of the Limitation of Size Conception. For instance, Scott (1974) uses them to derive Infinity and Replacement in his axiomatization of the Iterative Conception.[56] Although it is true that his reflection is first-order while Burgess' crucially is second-order (i.e., plural), it is still possible to envisage a plural-reflection principle *à la* Scott based on the relativization of a formula to a plural-level. This would seemingly preserve a certain amount of the "from below" spirit although the case for reflection principles within the Iterative Conception, in my opinion, remains controversial and in need of further clarification (see fn. 56). Moreover, the adoption of these principles should still be taken with caution. On one hand, Linnebo (2007) shows how, from assumptions that Burgess accepts in his use of reflection, one can easily derive that every plurality forms a set, which is inconsistent with the core assumption of P-COMP, as we explained in the opening. On the other hand, third-order reflection has been proven to be inconsistent[57] and therefore, assuming he could overcome Linnebo's challenge, Burgess may still fall prey of something like the Bad Company Problem for Neo-Fregeans.

Second, it is precisely a response to the Bad Company Problem that could reconcile the limitation of size component of Cantor's view with an iterative picture. Ironically, this theory should be labelled BBB (Boolos-Boolos-Boolos) as it is based on pivotal contributions to three areas: plural logic, the Iterative Conception and Neo-Fregeanism. In fact, Boolos (1987, 1989) retrieves Frege's program through a consistent reshaping of the infamous Basic Law V, called "New V", grounded on the notion of a *small concept*. This could capture Cantor's idea of consistent multiplicities, once the plural reading of second-order logic from Boolos (1984, 1985) has been implemented. Then, he contrasts this approach to the Iterative Conception (Boolos, 1971, 1989) in the usual terms of *logical* versus *combinatorial* collections (see §3.2). The former notion, he argues, has more traction to derive Separation, Replacement and Choice in their schematic or conceptual formulation. Since this fragment of ZFC is precisely what poses problems for the iterative-combinatorial picture, one may want to reconcile the two ideas after the pioneering work of Shapiro and Weir (1999). The two conceptions could work together to

[56] Scott was explicitly inspired by Lévy's seminal work (1960; 1961), while traces of the link between the conception and reflection can already be found in some remarks from Gödel (see Wang, 1977, p. 325, Wang, 1996, p. 285 and Koellner, 2009). However, one may also argue that Scott's appeal is illegitimate due to reflection principles clearly matching Uzquiano's "from above" characterization. Here I leave the question open, although I think that figuring out the status of reflection and related principles within the Iterative Conception is one of the most pressing issues in the philosophy of set theory, especially after the most recent developments of large cardinals and inner model programs. These seems to suggest an intrinsic link between the logical and the combinatorial component of the set concept, the former being more clearly represented by reflection. See in particular Bagaria and Ternullo (2025); Ternullo and Venturi (MS) who comment on this topic in connection to the most recent results in the search for new axioms (Aguilera et al., 2025; Aguilera, Bagaria, and Lücke, 2024).

[57] See Koellner (2003, 2009); Tait (1998, 2005).

smoothly retrieve the whole of ZFC from a fully Cantorian (plural-based) conception of set that keeps track of both the *logical-limitative* and of the *combinatorial-iterative* components.[58]

6.2. Potentialism and Critical Plural Logic

A second approach that must be mentioned is the one that pursues a modal understanding of Cantor's remarks, as in Øystein Linnebo's project of set-theoretic potentialism (2010; 2013).[59] Going back to Cantor's initial quotes, a perspicuous way of making sense of them is by the slogan "*set-existence* is a matter of *co-existence*" (Roberts, MS). However, this in not enough and the notion of co-existence needs a further sharpening: after all, under the common reading of plural logic, all the sets can co-exist as a plurality, but we don't want them to collapse into a set. In the recent debate a popular sharpening has been provided through a modal analysis that takes Cantor's use of modal expression at face value and re-interprets the slogan as: "*possible*" set-existence is a matter of "*possible*" co-existence.

The purported advantage of this approach, according to Roberts (MS), is that it does not modify the notion of co-existence, but rather takes the notion as it is and analyzes it in a modal context to make sense of it. This analysis ultimately yields a restriction of P-COMP, since its modal translation

$$\Diamond\exists xx\Box\forall x(x \prec xx \leftrightarrow \phi(x)) \qquad \text{(P-COMP}^{\Diamond}\text{)}$$

is not true for all conditions 'ϕ' but only for those that are "*extensionally definite*" (Linnebo, 2010, p. 157). Therefore, contemporary versions of modal set theories Sutto (2024), disqualify a priori pluralities such as those of all sets, ordinals or cardinals. This is in stark contrast with PLT, whose endorsement of P-COMP makes it distinctively actualist. In this sense, PLT represents the opposite view on co-existence as framed by Roberts' slogan: *set-existence* is a matter of *co-existence* at a stage or, better, *at a plural level*.[60]

If one finds the appeal to modalities problematic,[61] Critical Plural Logic (CPL) as advanced by Florio and Linnebo (2021) provides a non-modal theory grounded on the same idea of limiting P-COMP to make all multiplicities consistent so to speak. More precisely, Florio and Linnebo black-box the notion of "extensional definiteness" (i.e., consistency in Cantor's informal terms), which can instead be explicitly defined by potentialism, and axiomatize it. Their approach interestingly resembles Zermelo's original axiomatization of set theory if interpreted as black-boxing a consistent notion of collection and axiomatizing it.

No matter how one frames it, an issue with this approach is that it seems to be in stark contrast with Cantor's opening passages where he does not deny that inconsistent multiplicities are somehow conceivable or that they exist. What is outside the scope of "mathematical contemplation", in the passage from the letter to Hilbert, are the sets obtained from those pluralities, which are straight-away contradictory. The multiplicities themselves, on the contrary, are explicitly said to be "*definite*" (Cantor, 1899), a position that seems to put Cantor in agreement with unrestricted P-COMP. But what about co-existence then? If one forces

[58] This would also agree with the direction that contemporary set theory seems to be taking in some of its latest developments mentioned at the end of fn. 56. However, it would also pose a question on the status of pluralities *qua* combinatorial or logical collections, since this operation seems to require them to instantiate both features. The situation is similar to the one outlined by Maddy (1988) concerning her previous 1983 work on classes.

[59] The other main proponent of potentialism is Studd (2013, 2019). However, since he does not place Cantor and plural logic at the center of his investigation, here I focus on Linnebo's work.

[60] Under this respect I do not agree with Roberts (MS) in opposing the Iterative Conception to both potentialism and the limitation of size view, but I rather interpret potentialism as a modal approach to the conception in accordance with Sutto (2024).

[61] See Sutto (2024) for a survey on issues concerning modal approaches to potentialism and Button (MS) for a very careful critique.

the debate in terms of this notion, which may not be accepted by those who endorse a limitation of size view,[62] there seems to be no better way than framing the idea as *co-existence at a plural level* as articulated by the Iterative Conception *qua* instantiated by PLT.

However, despite Yablo's (2006) observation that favoring P-COMP is the route generally taken when the conflict with COLLAPSE is outlined, a fully articulated theory of sets in agreement with the former principle had yet to be explored. The reason is that most axiomatizations of the Iterative Conception are first-order and completely overlook the Cantorian conception of the process of set construction as plural-to-set abstraction. On top of that, Florio and Linnebo seem to further deny that such an account is philosophically defensible:

> We have described two very attractive applications of plural logic: as a way of giving an account of sets, and as a way of obtaining proper classes "for free". Regrettably, it looks like the two applications are incompatible. [...] Is there any way to retain both of the attractive applications of plural logic? To do so, we would have to restrict the domain of application of the "set of " operation so that the operation is undefined on the very large pluralities that correspond to proper classes, while it remains defined on smaller pluralities. The obvious concern is that this restriction would be *ad hoc*. (Florio and Linnebo, 2021, p. 72)

Against this, PLT shows that the two applications are compatible after all. Moreover, it does so in a non *ad hoc* way as the restriction on '$\ltimes$' is motivated by the Iterative Conception of Set, an intuitive and natural way of describing the process of set formation. Therefore, unless one wants do deny the naturalness of the conception, PLT seems to constitute a natural reply to the challenge posed by Florio and Linnebo. Remarkably, arguing against the conception is not an option for them, since they too seek inspiration in the famous Gödel's passage with which we started:

> To respond to this challenge, we might seek inspiration from Gödel, who points to a restriction when he requires that the "set of " operation be applied to "well-defined objects". [...] One option is to understand Gödel as requiring that the objects in question be properly circumscribed. [...] there are indeed "collections" that fail to be properly circumscribed. However, we also argue that every plurality is (in the appropriate sense) properly circumscribed and can thus figure as an argument of the "set of"operation. (Florio and Linnebo, 2021, p. 72)

As shown by PLT, a natural way to make sense of Gödel's claim is by interpreting "properly circumscribed" as "level-bound", in agreement with the developments of the Iterative Conception that stems precisely by that passage. Making all pluralities properly circumscribed seems to be an additional, rather substantial, assumption tied to an implicit potentialist understanding of the process of set formation. While challenging this approach is beyond the scope of this paper, I think that the contrast between PLT and CPL highlights an *interesting tension between two ways of interpreting the Cantorian plural conception of set and that ultimately boils down to the conflict between actualism and potentialism*. On my end, I argued that PLT does justice to an approach that seems to be more in line with Cantor's remarks, but that has not been hugely

[62] Co-existence as smallness does not seem a perspicuous explanation.

debated for various reasons, such as the focus on first-order axiomatizations of the Iterative Conception, the development of potentialism itself and the connection between plural-based axiomatization of sets and the Limitation of Size Conception as proposed by Burgess (2004); Pollard (1996). Of course this is far from settling the dispute, but at least the kind of actualism more in line with the Iterative Conception rather than the Limitation of Size view has now been provided with a carefully articulated theory that speaks the same plural-based language of potentialism.

6.3. Cantorian Set Theory

Another reason why an approach like PLT struggled to emerge, I think, is due to the only development of a plural iterative conception before mine being the one advanced by Oliver and Smiley (2016, 2018). More precisely, the fact that they explicitly place their Cantorian theory in opposition to standard axiomatic approaches like ZF and ZFC obscured the fact that such an account could also serve the purposes highlighted in this paper: making sense of Cantor plural remarks while being faithful to a conception of set tied to standard axiomatizations.[63] Moreover, since PLT could in principle be interpreted as a development of their own view, I argue that it also provides substantial improvements that go beyond the fact that it can derive pure axiomatic set theory. That is, PLT would perform better even if it was made to serve Oliver and Smiley's critique of the empty and singleton sets.[64] In particular, I think that they missed on three occasions to make their plural account of the conception "truly plural" so to speak.

First, despite the fact that they open the chapter on the theory by observing that "a great deal of reference to sets is merely an unnecessary and obfuscatory way of speaking [...] fuelled by the singularist drive to replace plural language by talk about sets" (2016, p. 245), their version of '¶' outputs sets rather than pluralities. That is, while their histories are pluralities, their potentiation applies to a plurality and produces a set. Therefore, their levels (still potentiations of histories) end up being sets rather than pluralities. I think that this is a substantial missed opportunity on having a completely plural characterization of the hierarchy of sets as PLT has. For instance, as noted above, the fact that our levels are pluralities allows to interpret PLT as a realization of a theory of levels with 1 level of classes which recaptures MK as speculated by Montague, Scott, and Tarski (unpublished). This is obviously something that is out of reach for a theory with only set-levels.

Second, rather than a primitive relation like my (and Burgess') '$\ltimes$', they start with the functional term '{}' which is the same as my '$\uparrow$'. While this also captures Gödel's "set of" operation, their endorsement of full P-COMP makes it denote a partial function, but they offer no explanation of when it may be total and are instead forced to appeal to a free logic with existence predicates both for plurals and for singular entities. That is, they do not realize they can use the level-theoretic setting to make sense of Cantor's partition between consistent and inconsistent multiplicities and also preserve classicality. To do so, however, appealing to a relational symbol is quite crucial, since '$\ltimes$' does not prejudge, as '$\uparrow$' does, whether there is such a thing as the collapsed set. This, I argue, is another missed opportunity: since they care so much for being faithful to Cantor why missing the chance of making sense of one of its most relevant yet obscure distinctions?

[63] This despite the fact that they too argued for a use of their theory to sanction ZF as I did. See (Oliver and Smiley, 2016, §14.8).

[64] This on top of the fact that my theory is based on an updated version of the theory they use as a blueprint, namely Potter (2004). Of course it would not be fair to accuse them of missing out a theory that was not even there when they developed their own. Nonetheless, the same improvements that LT brought to Potter's theory can also be appreciated for PLT with respect to their theory.

Third, their axioms seem to also miss some relevant points. For one thing, they postulate that a set is not included among the plurality which generates it, a fact easily derived from rules on how to alternate between '$\uparrow$' and '$\downarrow$' and the pivotal proof of well-ordering. Since they also aim for this proof, and given its importance in the context of the conception, this gap in their axiomatization is quite significant. Furthermore they assume as an axiom that sets uniquely determine their elements: $\uparrow xx = \uparrow yy \rightarrow xx \approx yy$. This is the opposite of Extensionality which, they claim, is implicit in the syntactic characterization of '$\{\}$' as a function. While this is true, it also completely reverses the order of explanation, which should go from the elements of a set to the set, not the other way around, betraying the Cantorian spirit of the project. Rather than have it implicit, making the other direction explicit and assuming as an axiom the biconditional, namely the Plural Law V from Florio and Linnebo (2021), would have eased the understanding of the process of set formation.

Therefore, despite the two projects come from different backgrounds, I think there are reason to favor PLT even to frame Oliver and Smiley's Cantorian set theory.

7. Conclusion

When introducing the Iterative Conception of Set, Dana Scott starts from a question concerning the Axiom of Separation: "where does the a [to which we apply Separation] come from?" (Scott, 1974, p. 208). His answer is of course the iterative process of set-formation based on a prototypical theory of stages or levels. An alternative reply, based on a Cantorian conception of set would be: Scott's a is collapsed from a given plurality aa. However, the question immediately resurfaces: *where does the plurality* aa *come from*? In this paper I proposed a reply that lines with Scott's: I described an iterative process of plural-to-set abstraction where pluralities collapse into sets as they appear at some level of a plural cumulative hierarchy of sets. To do so I resorted to the axiomatization of the conception in terms of a theory of plural levels, where the notion of a level is defined in explicit and non-recursive terms. The resulting Plural Level Theory yields a perspicuous explanation of Cantor's idea concerning the process of set-formation: *consistent* and *inconsistent multiplicities*, respectively, are explained as *level-bounded* and *level-unbounded pluralities*.

Besides reconciling the idea that sets are collapsed from a given plurality with the most popular conception of set, PLT also exhibits some nice results in line with the literature that inspires it. The most important are the well-ordering of the levels and the derivation of the standard axioms of Zermelo-Fraenkel set theory, which mean that PLT actually pins down the "plural skeleton" of the Cumulative Hierarchy of sets. Moreover, if one finds the language of plurals too exotic, PLT can be traced down to an equivalent nice theory of sets, Button's LT, which is definitionally equivalent to PLT modulo Boolos' plural reading of second-order quantification. The same reading enables the plural understanding of classes advocated by Uzquiano, which means that PLT can be made equivalent to (two-sorted) Morse-Kelley class theory. This is a crucial point because it permits an account of both notions of Cantorian multiplicities in terms of pluralities while maintaining, at the same time, a fundamental and non ad hoc link between some pluralities and the sets.

Finally, I argued that PLT performs better than other approaches when it comes to regimenting the idea that sets are obtained from a process of plural-abstraction. First, although some may argue that Cantor's original view squares better with a limitation of size approach, I argued that PLT offers a more satisfactory account of Cantor's notion of multiplicity than

Burgess' BB. The best chance to make a limitation of size approach work, I also argued, is to put it side by side with an iterative view to take care of both the combinatorial and the logical aspect of the process of set formation. Second, I showed how PLT meets the challenge advanced by potentialism, which argues that a plural account of sets grounded on full comprehension is not philosophically justifiable. On the contrary, PLT not only shows that this is possible, but does so by departing from the same observation on the Iterative Conception that the non-modal account of potentialism favors. Third, and finally, I showed how PLT performs better than the analogous project developed by Oliver and Smiley. Notably this is true even if we set aside the dispute over the status of singletons and the empty set and focus solely on the core idea of pluralities as the grounds for the process of set formation.

Of course many more questions concerning PLT and its rivals remain unanswered. In particular, the conflict between actualism and potentialism, on one side, and the one between the Iterative Conception and Limitation of Size, on the other, represent pivotal crossroads in the philosophy of set theory. While it is beyond the scope of this paper to settle them, here I sketched some of the possible replies provided by PLT. My overall aim was to present and do justice to a plural-based account of the process of set-formation that, setting aside Oliver and Smiley non-standard set theory, was still missing from a literature that mostly focused on Limitation of Size (between the late 1990s and the early 2000s) and on potentialism (in the past fifteen years). Now that a plural iterative conception of set has been carefully outlined, the stage is set for a fair debate on which approach best captures the intuitions of the father of set theory.

A. Appendix

1.1. The Proof of Well-ordering

First of all, remember that the proof goes through simply by appeal to Plural Separation rather than full Comprehension. Moreover, to avoid specifying it every time, we assume that all the lemmas where we apply P-COLLAPSE below are conditionals on the levels being bounded. This can be done without loss of generality since, in general, all levels except the last one, namely the first class-level, are always bounded. In other terms, this is like saying that the well-ordering happens within the last level, which is not different from a general statement of well-ordering for a class.

Let's start with some basic facts and definitions. First, (4) and (5) from Def. 2.3 yield plural *transitivity* and *super-transitivity*, granted an analogue of (5) for sets after Button (2024): '$a \rhd b \leftrightarrow_{def} (\exists c \in b)a \subseteq c$':

Definition A.1. A set a is TRANSITIVE iff $(\forall x \in a)x \subseteq a$. A plurality aa is TRANSITIVE iff $(\forall x \prec aa)x \sqsubseteq aa$.

Definition A.2. A set a is SUPER-TRANSITIVE iff $(\forall x \rhd a)x \in a$.[65] A plurality aa is SUPER-TRANSITIVE iff $(\forall x \blacktriangleright aa)x \prec aa$.

A significant fact is then that transitivity and super-transitivity carry over from the pluralities to the respective collapsed sets:

Fact A.3. Any set collapsed from a (super)-transitive plurality is (super)-transitive.

[65] Button (2021, fn. 10) notes that the property of super-transitivity (for sets) has many different names in the literature. He chooses the label "potent" to highlight the connection with potentiation. Here I prefer to stick to "super-transitive", also adopted by Linnebo (2007), to better highlight the connection with the more familiar notion of transitivity.

Proof. Consider an arbitrary transitive and super-transitive aa and assume that $\uparrow aa$ exists. TRANSITIVITY: consider an arbitrary $x \in \uparrow aa$ and an arbitrary $y \in x$; since $x \in \uparrow aa$, $x \prec aa$ (Def. 2.3.1); therefore $y \in x \prec aa$, but aa is transitive, so $y \prec aa$ and, again by Def. 2.3.1, $y \in \uparrow aa$. SUPER-TRANSITIVITY: consider an arbitrary x and let c be a set such that $x \subseteq c \wedge c \in \uparrow aa$; by Def. 2.3.1 $c \prec aa$; therefore, $x \subseteq c \prec aa$, but aa is super-transitive so $x \prec aa$ and, by Def. 2.3.1, $x \in a$. □

The following fact about plural potentiation is trivial but worth mentioning:

Fact A.4. $uu \preccurlyeq \P\P(uu)$.

Proof. Consider an arbitrary $x \prec uu$. To have $x \prec \P\P(uu)$ means that there is a c such that $x \subseteq c \prec uu$. The fact trivially follows by instantiating c with x. □

We can then trace a connection between the notion of plural super-transitivity and of plural potentiation:

Lemma A.5. *$\P\P(aa)$ is super-transitive.*

Proof. Assume that $\P\P(aa)$ exists and consider an x such that $x \blacktriangleright \P\P(aa)$. So there is a $c \prec \P\P(aa)$ such that $x \subseteq c$, that is, $c \blacktriangleright aa$. But again, so there is a $b \prec aa$ such that $c \subseteq b$. Since $x \subseteq c \subseteq b$, we can conclude $x \prec \P\P(aa)$. □

Lemma A.6. *aa is super-transitive iff $aa \approx \P\P(aa)$.*

Proof. $\Leftarrow$ follows from Lemma A.5. $\Rightarrow$: assume aa is super transitive and $\P\P(aa)$ exists. Therefore, $x \blacktriangleright aa$ and so $x \prec aa$ by super-transitivity and Fact A.4. □

Lemma A.7. *Every level is transitive and super-transitive.*

Proof. Fix a level $ss \approx \P\P(uu)$, for some history uu. TRANSITIVITY: consider $a \prec ss$ and $x \in a$. Since $a \prec ss \approx \P\P(uu)$, $a \subseteq c \prec uu$ for some c. Since $x \in a$ and $a \subseteq c$, then $x \in c$. From Def. 4.4 $c = \uparrow(\P\P(c \cap uu))$, therefore, $x \prec \P\P(c \cap uu)$, but $\P\P(c \cap uu) \preccurlyeq \P\P(uu)$, so $x \prec \P\P(uu) \approx ss$. SUPER-TRANSITIVITY: use Lemma A.6. □

Up to now everything matches the proof in Button (2021). Before proving minimality, as LT does, some results take care of the type differences through '$\ltimes$'.[66]

Lemma A.8. *If every Φ is super-transitive, some uu are Φ and there is a level ss such that $uu \not\preccurlyeq ss$, then there is some tt which is a $\ltimes$-minimal Φ: $\Phi(tt) \wedge \forall aa(\Phi(aa) \rightarrow \uparrow aa \not\prec tt)$.*

[66] Special thanks to Tim Button for pointing me out this crucial passage, without which the proof remained stuck for months.

Proof. Assume some uu such that $\Phi(uu)$ and assume some level ss such that $uu \preccurlyeq ss$. Apply P-SEP twice:

$$cc :\approx \|x \prec uu : \forall ss((\Phi(ss) \wedge \textsc{Lev}(ss)) \rightarrow x \prec ss)\| \approx$$
$$\approx \|x : \forall ss((\Phi(ss) \wedge \textsc{Lev}(ss)) \rightarrow x \prec ss)\|$$
$$dd :\approx \|x \prec cc : x \notin x\|$$

Note that $dd \preccurlyeq cc \preccurlyeq uu \preccurlyeq ss$ and since $\textsc{Lev}(ss)$ we can apply P-COLLAPSE to obtain $d =\uparrow dd$. Of course $d \not\prec cc$, otherwise $d \in d \leftrightarrow d \prec dd \leftrightarrow d \notin d$, which is absurd. So there must be some tt such that $\Phi(tt)$ and $d \not\prec tt$. For reductio assume there is a vv such that $\Phi(vv)$ and $\uparrow vv \prec tt$. Since $d \subseteq \uparrow vv$ and since every Φ is super-transitive we have $d \prec tt$, contradiction. □

Lemma A.9. *If hh is a plural history such that $hh \preccurlyeq ss$ for some level ss and $a \prec hh$, then $\downarrow a$ is a level.*

Proof. For reductio assume there is a history hh and a level ss such that $hh \preccurlyeq ss$, but that the conclusion of the lemma does not follow. Before moving on note the following corollary:

Corollary A.10. *If $a \prec hh$, then $\downarrow a$ exists.*

Proof. Assume $a \prec hh$. Since $hh \preccurlyeq ss$, $a \prec ss$ and since every level is transitive $a \sqsubseteq ss$. Apply P-SEP: $aa \approx \|x : x \prec ss \wedge x \in a\| \approx \|x : x \in a\| \approx \downarrow a$. □

Therefore we can now apply Lemma A.8: fix some $a \prec hh$ such that $\downarrow a$ is a $\ltimes$-minimal non-level, i.e., $(b \prec hh \wedge \textsc{Lev}(\downarrow b)) \rightarrow b \not\prec \downarrow a$. This can also be rephrased as $\forall b \prec hh(b \in a \rightarrow \textsc{Lev}(\downarrow b))$. Moreover, by $a \prec hh$ we know that $a =\uparrow (\P\P(a \cap hh))$, thus if we show that $(a \cap hh)$ is a history, $\downarrow a$ will be a level, contradiction. Assume a $b \prec (a \cap hh)$, that is, $b \in a \wedge b \prec hh$. Since $b \in a$, $\downarrow b$ is a level and, since $b \prec hh$, $b =\uparrow (\P\P(b \cap hh))$. Fix some $x \in b$, that is, $x \prec \downarrow b$, but levels are transitive so $x \sqsubseteq \downarrow b$, hence $x \subseteq b$. Then, since $b \in a$, i.e., $b \prec \downarrow a$, we have $x \blacktriangleright \downarrow a$ as $\downarrow a$ is super-transitive ($a \prec hh$ + Corollary A.10 + Lemma A.3), and so $x \in a$. Therefore $b \subseteq a$, so $b =\uparrow (\P\P(b \cap hh)) =\uparrow (\P\P(b \cap (a \cap hh)))$, so $(a \cap hh)$ is a history. □

Lemma A.11. *If ss is a level, then $ss \approx \P\P(rr)$, with $rr \approx \|t : t \prec ss \wedge \textsc{Lev}(\downarrow t)\|$.*

Proof. ($\Leftarrow$). Assume ss is a level and $a \prec \P\P(rr)$, that is, $a \blacktriangleright rr$. So, from Def. 2.3 and the definition of rr we know that there is some $t \supseteq a$ with $t \prec rr$, that is, $t \prec ss$. But then $a \blacktriangleright ss$ and since levels are super-transitive $a \prec ss$. ($\Rightarrow$). Assume $a \prec ss$. By Def. 4.5 we know that there is a history hh such that $ss \approx \P\P(hh)$ and so $a \prec \P\P(hh)$. This means that $a \blacktriangleright hh$, that is, there is a $t \prec hh$ with $t \supseteq a$. Remember that by Fact A.4 $hh \preccurlyeq \P\P(hh)$ and so $hh \preccurlyeq ss$, which allows the use of Lemma A.9 to say that $\downarrow t$ is a level. This plus $t \prec hh \preccurlyeq ss$ yield $t \prec rr$. □

Lemma A.12. *Levels are $\ltimes$-comparable:*

$$\forall ss \forall tt((\textsc{Lev}(ss) \wedge \textsc{Lev}(tt)) \rightarrow (\uparrow ss \prec tt \vee ss \approx tt \vee \uparrow tt \prec ss))$$

Proof. Suppose, for reductio, that some levels are $\ltimes$-incomparable. Since Lemma A.8 applies trivially to levels, we can assume a $\ltimes$-minimal level ss which is $\ltimes$-incomparable with some level. This means that for any $r \prec ss$ with $\textsc{Lev}(\downarrow r)$, $\downarrow r$ is $\ltimes$-comparable with all levels. Another round of Lemma A.8 provides us with another $\ltimes$-minimal level tt which is $\ltimes$-incomparable with ss. I shall show that $ss \approx tt$, contradiction. Fix some $a \prec ss$. By Lemma A.11 there is some $r \prec ss$ with $\textsc{Lev}(\downarrow r)$ and $a \subseteq r$. Since $\downarrow r$ is $\ltimes$-comparable with all levels this is also true for tt. But, if $\downarrow r \approx tt$, then $r = \uparrow tt$ and so $\uparrow tt \prec ss$, contradicting choice of tt. On the other hand, if $\uparrow tt \prec \downarrow r$, since levels are transitive we'd have $\uparrow tt \prec ss$, contradicting again choice of tt. So it must be that $r \prec tt$. Since levels are super-transitive we have $a \prec tt$ and thus $ss \preccurlyeq tt$. A similar reasoning yields $tt \preccurlyeq ss$, hence $ss \approx tt$. □

Overall, Lemma A.8 and A.12 tell us that $\ltimes$ acts as a sort of "trans-type well-ordering".

Lemma A.13. *If some plural level is Φ, then there is a $\preccurlyeq$-minimal level which is Φ. Formally:* $\exists rr(\textsc{Lev}(rr) \wedge \Phi(rr)) \rightarrow \exists ss((\textsc{Lev}(ss) \wedge \Phi(ss)) \wedge \forall rr((\textsc{Lev}(rr) \wedge \Phi(rr)) \rightarrow (rr \preccurlyeq ss \rightarrow rr \approx ss)))$.

Proof. As Lemma A.8 trivially applies to levels, repeat its steps to obtain a $\ltimes$-minimal plural level ss. Assume there is a plural level tt such that $\Phi(tt)$ and $tt \preccurlyeq ss$. Since $\uparrow tt \not\prec ss$, by Lemma A.12 either $ss \approx tt$, in which case we are done, or $\uparrow ss \prec tt$. But levels are transitive so $\uparrow ss \sqsubseteq tt$, hence $ss \preccurlyeq tt$. □

In general, this lemma tells us that $\ltimes$-minimality implies $\preccurlyeq$-minimality.[67] While the former is, in a sense, more fundamental, we still need the latter to properly state the usual "intra-type" kind of well-ordering.

Lemma A.14. *All levels are comparable:*

$$\forall ss \forall tt((\textsc{Lev}(ss) \wedge \textsc{Lev}(tt)) \rightarrow (ss \preccurlyeq tt \vee ss \approx tt \vee tt \preccurlyeq ss)).$$

Proof. Follow the steps of Lemma A.12 substituting $\ltimes$-minimal with $\preccurlyeq$-minimal. □

We are now ready to state the fundamental theorem of Plural Level Theory:

Theorem A.15. *The levels are well-ordered by plural inclusion ($\preccurlyeq$).*

Proof. The theorem follows from Lemma A.13 together with Lemma A.14; □

As for LT, Lemmas A.8 and A.12, together with P-STRAT, allow us to consider the level at which the elements of a set first appear.

Definition A.16. If $\uparrow aa$ exists, ℓaa is the $\ltimes$-least level that contains aa: $aa \preccurlyeq \ell aa$ and $\forall ss((\textsc{Lev}(ss) \wedge aa \preccurlyeq ss) \rightarrow \uparrow ss \not\prec \ell aa)$.

[67] For reasons of type bookkeeping, namely the fact that we are moving between higher and lower types, the other direction seems to not obviously follow. While it is not an issue for the overall theory, I leave the question open for further investigations since it may provide interesting insights on these type raising/lowering phenomena.

This is a rather powerful tool since, together with the above results, it sanctions $\ltimes$-induction, which ultimately yields to the familiar $\in$-induction. This is also how we prove that no set is among the plurality from which it collapses, a feature that Oliver and Smiley missed.[68] Moreover, it also yields some intuitive properties of the levels, analogous to Button (2021, Lemma 3.12): e.g., $\uparrow aa \not\prec \ell aa$ parallels LT's 3.12(2) and expresses the "priority" of the elements of a set to the set itself.[69]

1.2. Bounded Pluralities

Proposition 5.5, PLT is enough to sanction the Kuratowski notation for ordered pairs. We can then use P-COMP to define "relational pluralities" as pluralities of ordered pairs. In fact, we can do more and define "functional pluralities" as relational pluralities where no two ordered pairs share the first element. Finally, we can relativize this notion to a given set to define the plurality of functions on that set:

Definition A.17 (FUNCTIONAL PLURALITY)**.** Given a set a, we define a FUNCTIONAL PLURALITY ON a as

$$ff^a :\approx \|x : \forall y \in a(\exists z(x = \langle y, z\rangle) \wedge \forall x, x', z, z'((x = \langle y, z\rangle \wedge x' = \langle y, z'\rangle) \rightarrow z = z'))\|$$

We can now define the "plural image" of a functional plurality:

Definition A.18 (PLURAL IMAGE)**.** Given a set a and the functional plurality ff^a on a, the PLURAL IMAGE OF ff^a is $ff^{[a]} :\approx \|y : (\exists x \in a)(\exists p \prec ff^a)p = \langle x, y\rangle\|$.

Remember that all these pluralities exist simply by P-COMP and since in $\mathcal{L}_{\prec,\ltimes}$ we are availing ourselves of unrestricted plural quantification we can freely quantify over them to obtain the final principle.

Acknowledgements. I would like to thank Øystein Linnebo and Salvatore Florio for their extensive feedbacks. The paper greatly benefited from two research stays at the University College London and at the University of Southern California. My most heartfelt thanks to Tim Button (UCL), who pulled me out of the swamp of the proof of well-ordering and provided helpful insights on the level-theoretic framework, and to Gabriel Uzquiano (USC), with whom I discussed the manuscript at length and with passion until the paper reached its final shape. Thanks also to Andrew Bacon, Neil Barton, Clara Bortoletto, Antonio Maria Cleani, Pablo Dopico, Joel David Hamkins, Matteo Plebani, Sam Roberts, Chris Scambler, Nicolò Siviero, Giorgio Venturi and Brandon Ward for further helpful comments and discussions, as well as to audiences in Konstanz, London (KCL), Los Angeles (USC), New York (CUNY), Oslo (UiO), Pisa (UniPi), Rome (Tor Vergata) and Vercelli (UPO). At some of these places earlier versions were presented under the title "Plural Level Theory" and "The Iterative Conception of Pluralities".

Funding. This research was supported by the European Union (ERC Advanced Grant, C-FORS: Construction in the Formal Sciences, awarded to Øystein Linnebo. Project number: 101054836). The research stay at the University of Southern California was further supported by a travel grant for longer research stays awarded by the University of Oslo, Faculty of Humanities (ref: 2023/13635).

[68] I leave this as an exercise to the reader.

[69] I redirect the reader to Button's paper for the proofs of these features since they match almost perfectly and are not so interesting.

References

Aguilera, J. P., Bagaria, J., Goldberg, G., and Lücke, P. (2025). Large cardinals beyond HOD. *arXiv*. https://arxiv.org/abs/2509.10254.

Aguilera, J. P., Bagaria, J., and Lücke, P. (2024). Large cardinals, structural reflection, and the HOD Conjecture. *arXiv*. https://arxiv.org/abs/2411.11568.

Bagaria, J. and Ternullo, C. (2025). Intrinsic Justification for Large Cardinals and Structural Reflection. *Philosophia Mathematica*, 33(2):123–154. DOI: https://doi.org/10.1093/philmat/nkaf006.

Barton, N. (2024). *Iterative Conceptions of Sets. Cambridge Elements: The Philosophy of Mathematics*. Cambridge University Press. DOI: https://doi.org/10.1017/978100922722.

Boolos, G. (1971). The Iterative Conception of Set. *Journal of Philosophy*, 68(8):215–231. Reprinted in Boolos (1998), pp. 13–29. DOI: https://doi.org/10.2307/2025204.

Boolos, G. (1984a). The Justification of Mathematical Induction. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1984(2):469–475. Reprinted in Boolos (1998, pp. 370–375). DOI: https://doi.org/10.1086/psaprocbienmeetp.1984.2.192521.

Boolos, G. (1984b). To Be is to Be a Value of a Variable (or to Be Some Values of Some Variables). *Journal of Philosophy*, 81(8):430–449. Reprinted in Boolos (1998, pp. 54–72). DOI: https://doi.org/jphil198481840.

Boolos, G. (1985). Nominalist Platonism. *Philosophical Review*, 94(3):327–344. Reprinted in Boolos (1998, pp. 73–87). DOI: https://doi.org/10.2307/2185003.

Boolos, G. (1987). IX - Saving Frege From Contradiction. *Proceedings of the Aristotelian Society*, 87(1):137–152. Reprinted in Boolos (1998, pp. 171–182). DOI: https://doi.org/10.1093/aristotelian/87.1.137.

Boolos, G. (1989). Iteration Again. *Philosophical Topics*, 17(2):5–21. Reprinted in Boolos (1998), pp. 88–104. DOI: https://doi.org/10.5840/philtopics19891721.

Boolos, G. (1998). *Logic, Logic, and Logic*. Harvard University Press.

Burgess, J. P. (2004). E Pluribus Unum: Plural Logic and Set Theory. *Philosophia Mathematica*, 12(3):193–221. DOI: https://doi.org/10.1093/philmat/12.3.193.

Button, T. (2021). Level Theory, Part 1: Axiomatizing the Bare Idea of a Cumulative Hierarchy of Sets. *Bulletin of Symbolic Logic*, 27(4):436–460. DOI: https://doi.org/10.1017/bsl.2021.13.

Button, T. (2024). Wand/Set Theories: A Realization of Conway's Mathematicians' Liberation Movement, with an Application to Church's Set Theory with a Universal Set. *Journal of Symbolic Logic*, pages 1–40. DOI: https://doi.org/10.1017/jsl.2024.21.

Button, T. (MS). Why Are All the Sets All the Sets? Manuscript.

Button, T. and Walsh, S. (2018). *Philosophy and Model Theory*. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198790396.001.0001.

Cantor, G. (1883). *Grundlagen einer allgemeinen Mannigfaltigkeitslehre. Ein mathematisch philosophischer Versuch in der Lehre des Unendlichen*. Leipzig. Translated in Ewald (1996, pp. 878–920). DOI: https://doi.org/10.1093/oso/9780198505365.003.0004.

Cantor, G. (1895). Beiträge zur Begründung der transfiniten Mengenlehre. *Mathematische Annalen*, 46(4):481–512. Translated in Cantor (1955, pp. 85–136). DOI: https://doi.org/10.1007/BF02124929.

Cantor, G. (1897). Letter to Hilbert 2 October 1897. Translated in Ewald (1996, pp. 927–928). DOI: https://doi.org/10.1093/oso/9780198505365.003.0004.

Cantor, G. (1899). Letter to Dedekind 3 August 1899. Translated in Ewald (1996, pp. 931–935). DOI: https://doi.org/10.1093/oso/9780198505365.003.0004.

Cantor, G. (1955). *Contributions to the Founding of the Theory of Transfinite Numbers*. Doover Publications, Inc. Translated by Philip E. B. Jourdain. Unabridged and unaltered reprint of the English translation first published in 1915.

Ewald, W. B., editor (1996). *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*, volume II. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198505365.001.0001.

Ferreirós, J. (2007). *Labyrinth of Thought. A History of Set Theory and Its Role in Modern Mathematics*. Birkhäuser. Second revised edition. DOI: https://doi.org/10.1007/978-3-7643-8350-3.

Florio, S. and Linnebo, O. (2021). *The Many and the One: A Philosophical Study of Plural Logic*. Oxford, England: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198791522.001.0001.

Forster, T. (2008). The Iterative Conception of Set. *Review of Symbolic Logic*, 1(1):97–110. DOI: https://doi.org/10.1017/s1755020308080064.

Fraenkel, A., Bar-Hillel, Y., and Lévy, A. (1973). *Foundations of Set Theory. Second Revised Edition*, volume 67 of *Studies in Logic and the Foundations of Mathematics*. Elsevier. DOI https://doi.org/10.1016/S0049-237X(08)70334-3.

Gödel, K. (1947). What is Cantor's Continuum Problem? *The American Mathematical Monthly*, 54(9):515–525. DOI: https://doi.org/10.2307/2304666.

Gödel, K. (1951). Some basic theorems on the foundation of mathematics and their implications. 25^{th} Josiah Willard Gibbs Lecture given at a meeting of the American Mathematical Society in Brown University, 26 December 1951. Reprinted in Gödel (1995, pp. 290–323). DOI: https://doi.org/10.1093/oso/9780195072556.003.0014.

Gödel, K. (1995). *Collected Works. Volume III. Unpublished Essays and Lectures*, volume III. Oxford, New York: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780195072556.001.0001.

Hallett, M. (1984). *Cantorian Set Theory and Limitation of Size*. Clarendon Press.

Incurvati, L. (2012). How to be a minimalist about sets. *Philosophical Studies*, 159(1):69–87. DOI: https://doi.org/10.1007/s11098-010-9690-1.

Incurvati, L. (2020). *Conceptions of Set and the Foundations of Mathematics*. Cambridge University Press. DOI: https://doi.org/10.1017/9781108596961.

Incurvati, L. (2025). Iteration and Dependence Again. In Antos, C., Barton, N., and Venturi, G., editors, *The Palgrave Companion to the Philosophy of Set Theory*, pages 247–271. DOI: https://doi.org/10.1007/978-3-031-62387-5_10.

Kanamori, A. (2004). Zermelo and Set Theory. *The Bulletin of Symbolic Logic*, 10(4):487–553. DOI: https://doi.org/10.2178/bsl/1102083759.

Klement, K. C. (2014). Early Russell on Types and Plurals. *Journal for the History of Analytical Philosophy*, 2(6):1–21. DOI: https://doi.org/10.15173/jhap.v2i6.47.

Koellner, P. (2003). *The Search for New Axioms*. PhD thesis, Massachusetts Institute of Technology.

Koellner, P. (2009). On Reflection Principles. *Annals of Pure and Applied Logic*, 157(2-3):206–219.

Kreisel, G. (1965). Mathematical Logic. In Saaty, T. L., editor, *Lectures in Modern Mathematics. Volume III*, pages 95–195. Wiley.

Kunen, K. (2013). *Set theory*. College Publications.

Landman, F. (1989). Groups, I. *Linguistics and Philosophy*, 12(5):559–605. DOI: https://doi.org/10.1007/bf00627774.

Lévy, A. (1960). Axiom Schemata of Strong Infinity in Axiomatic Set Theory. *Pacific Journal of Mathematics*, 10(1):223–238.

Lévy, A. and Vaught, R. L. (1961). Principles of partial reflection in the set theories of Zermelo and Ackermann. *Pacific Journal of Mathematics*, 11(3):1045–1062. DOI: http://dx.doi.org/10.2140/pjm.1961.11.1045.
Linnebo, O. (2007). Burgess on Plural Logic and Set Theory. *Philosophia Mathematica*, 15(1):79–93. DOI: https://doi.org/10.1093/philmat/nkl029.
Linnebo, O. (2010). Pluralities and Sets. *Journal of Philosophy*, 107(3):144–164. DOI: https://doi.org/10.5840/jphil2010107311.
Linnebo, O. (2013). The Potential Hierarchy of Sets. *Review of Symbolic Logic*, 6(2):205–228. DOI: https://doi.org/10.1017/s1755020313000014.
Linnebo, O. and Nicolas, D. (2008). Superplurals in English. *Analysis*, 68(3):186–197. DOI: https://doi.org/10.1093/analys/68.3.186.
Maddy, P. (1983). Proper Classes. *Journal of Symbolic Logic*, 48(1):113–139.
Maddy, P. (1988). Believing the Axioms. I. *Journal of Symbolic Logic*, 53(2):481–511. DOI: https://doi.org/10.2307/2274520.
Montague, R. (1965). Set Theory and Higher-Order Logic. In Crossley, J. N. and Dummett, M., editors, *Formal systems and recursive functions. Proceedings of the Eight Logic Colloquium, July 1963*, volume 40 of *Studies in Logic and the Foundations of Mathematics*, pages 131–148. North-Holland. DOI: https://doi.org/10.1016/S0049-237X(08)71686-0.
Montague, R., Scott, D., and Tarski, A. (unpublished). An Axiomatic Approach to Set Theory. Archive Copy from the Bancroft Library: Alfred Tarski Papers, circa 1923-1985 (BANC MSS 84/69 c, Carton 4, Folder 29-30).
Nicolas, D. and Payton, J. D. (2025). Superplurals Analyzed Away. *Inquiry*. DOI: https://doi.org/10.1080/0020174x.2024.2440765.
Oliver, A. and Smiley, T. (2016). *Plural Logic: Second Edition, Revised and Enlarged*. Oxford University Press.
Oliver, A. and Smiley, T. (2018). Cantorian Set Theory. *Bulletin of Symbolic Logic*, 24(4):393–451. DOI: https://doi.org/10.1017/bsl.2018.10.
Parsons, C. (1974). Sets and Classes. *Noûs*, 8(1):1–12. Reprinted in Parsons (1983), pp. 209–220. DOI: https://doi.org/10.2307/2214641.
Parsons, C. (1983). *Mathematics in Philosophy: Selected Essays*. Cornell University Press.
Payton, J. D. (2025). From Singular to Plural...And Beyond? *Philosophy and Phenomenological Research*, 110(3):983–1012. DOI: https://doi.org/10.1111/phpr.13136.
Pollard, S. (1985). Plural quantification and the iterative concept of set. *Philosophy Research Archives*, 11:579–587. DOI: https://doi.org/10.5840/pra19851134.
Pollard, S. (1996). Sets, Wholes and Limited Pluralities. *Philosophia Mathematica*, 4(1):42–58. DOI: https://doi.org/10.1093/philmat/4.1.42.
Potter, M. (1990). *Sets: An Introduction*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198533887.001.0001.
Potter, M. (2004). *Set Theory and its Philosophy: A Critical Introduction*. Oxford, England: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199269730.001.0001.
Roberts, S. (2022). Pluralities as Nothing Over and Above. *The Journal of Philosophy*, 119(8):405–424. DOI: https://doi.org/10.5840/jphil2022119828.
Roberts, S. (MS). Ultimate V. Manuscript.
Russell, B. (1903). *The Principles of Mathematics*. Cambridge University Press. DOI: https://doi.org/10.4324/9780203822586.
Scott, D. (1974). Axiomatizing Set Theory. In Jech, T. J., editor, *Axiomatic Set Theory. Proceedings of Symposia in Pure Mathematics. Volume XIII, Part II*, pages 207–214. American Mathematical Society, Providence, Rhode Island. DOI: https://doi.org/10.1090/pspum/013.2/0392570.

Shapiro, S. and Weir, A. (1999). New V, ZF and Abstraction. *Philosophia Mathematica*, 7(3):293–321. DOI: https://doi.org/10.1093/philmat/7.3.293.

Shoenfield, J. R. (1967). *Mathematical Logic*. Reading, MA, USA: Reading, Mass., Addison-Wesley Pub. Co. DOI: https://doi.org/10.1201/9780203749456.

Shoenfield, J. R. (1977). Axioms of Set Theory. In Barwise, J., editor, *Handbook of Mathematical Logic*, volume 90 of *Studies in Logic and the Foundations of Mathematics*, pages 321–344. North-Holland Pub. Co. DOI: https://doi.org/10.1016/S0049-237X(08)71106-6.

Studd, J. P. (2013). The Iterative Conception of Set: A (Bi-)Modal Axiomatisation. *Journal of Philosophical Logic*, 42(5):697–725. DOI: https://doi.org/10.1007/s10992-012-9245-3.

Studd, J. P. (2019). *Everything, More or Less: A Defence of Generality Relativism*. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198719649.001.0001.

Sutto, D. (2024). A Taxonomy for Set-Theoretic Potentialism. *Philosophia Mathematica*, pages 1–28. DOI: https://doi.org/10.1093/philmat/nkae016.

Tait, W. (1998). Zermelo's Conception of Set Theory and Reflection Principles. In Schirn, M., editor, *The Philosophy of Mathematics Today*. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198236542.003.0019.

Tait, W. (2005). Constructing Cardinals from Below. In *The Provenance of Pure Reason: Essays in the Philosophy of Mathematics and Its History*. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780195141924.003.0007.

Ternullo, C. and Venturi, G. (MS). Large Cardinals and the Inner Model Program. Manuscript.

Uzquiano, G. (2003). Plural Quantification and Classes. *Philosophia Mathematica*, 11(1):67–81. DOI: https://doi.org/10.1093/philmat/11.1.67.

van Heijenoort, J., editor (1967). *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931*. Harvard University Press.

von Neumann, J. (1925). Eine Axiomatisierung der Mengenlehre. *Journal für die reine und angewandte Mathematik*, (154):219–240. Translated as "An axiomatization of set theory" in van Heijenoort (1967, pp. 393–413).

von Neumann, J. (1928). Die Axiomatisierung der Mengenlehre. *Mathematische Zeitschrift*, 27:669–752. DOI: https://doi.org/10.1007/BF01171122.

Wang, H. (1974). *From Mathematics to Philosophy*. London and Boston: Routledge. DOI: https://doi.org/10.2307/2217640.

Wang, H. (1977). *Large Sets*, pages 309–333. Springer Netherlands, Dordrecht. DOI: https://doi.org/10.1007/978-94-010-1138-9_17.

Wang, H. (1996). *A Logical Journey: From Gödel to Philosophy*. Bradford. DOI: https://doi.org/10.7551/mitpress/4321.001.0001.

Yablo, S. (2006). Circularity and Paradox. In Bolander, T., Hendricks, V. F., and Pedersen, S. A., editors, *Self-Reference*, pages 139–157. CSLI Publications.

Zermelo, E. (1930). Über Grenzzahlen und Mengenbereiche. *Fundamenta Mathematicae*, 16:29–47. Translated in Ewald (1996), pp. 1219–1233. DOI: https://doi.org/10.4064/fm-16-1-29-47.